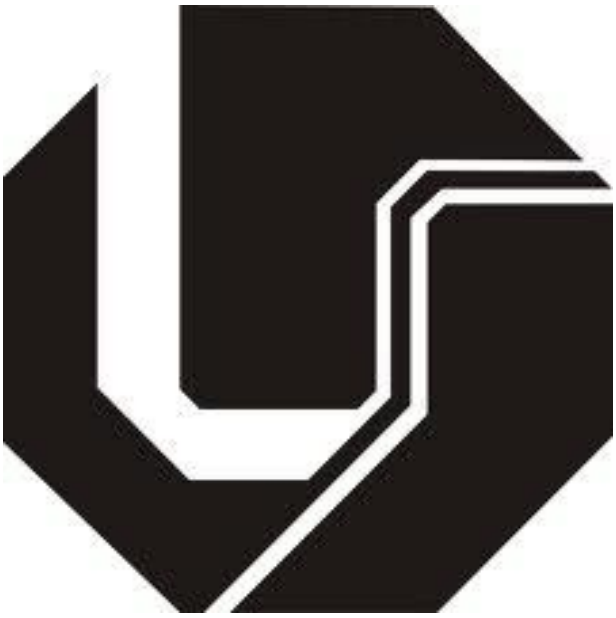
UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA
PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA



ANÁLISE DO ÍNDICE DE NEBULOSIDADE PARA OTIMIZAÇÃO DO PROCESSO DE AGRUPAMENTOS DE DADOS

ERNANI CLÁUDIO BORGES

UBERLÂNDIA
2012

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA
PÓS-GRADUAÇÃO EM ENGENHERIA ELÉTRICA

ANÁLISE DO ÍNDICE DE NEBULOSIDADE PARA OTIMIZAÇÃO DO PROCESSO DE AGRUPAMENTOS DE DADOS

Dissertação apresentada por Ernani Cláudio Borges à Universidade Federal de Uberlândia para obtenção do título de Mestre em Ciências aprovada em **09 de Outubro de 2012** pela Banca Examinadora:

Prof. Dr. Adriano Alves Pereira (UFU Orientador)
Prof. Dr. Adriano de Oliveira Andrade
Prof. PhD. Ailton Akira Shinoda

UBERLÂNDIA (MG)
2012

ANÁLISE DO ÍNDICE DE NEBULOSIDADE PARA OTIMIZAÇÃO DO PROCESSO DE AGRUPAMENTOS DE DADOS

ERNANI CLÁUDIO BORGES

Dissertação apresentada por Ernani Cláudio Borges à Universidade Federal de Uberlândia como parte dos requisitos para obtenção do título de Mestre em Ciências. Área de concentração: Processamento da Informação. Linha de Pesquisa: Processamento digital de sinais.

_____

Professor Adriano Alves Pereira, Dr.
Orientador

_____

Professora Selma Terezinha Milagre, Dra.
Co-Orientadora

_____

Professor Alexandre Cardoso, Dr.
Coordenador do Curso de Pós-Graduação

Uberlândia-MG, 2012

Dados Internacionais de Catalogação na Publicação (CIP)

Sistema de Bibliotecas da UFU, MG - Brasil

B732a 2012

Borges, Ernani Cláudio, 1971-

Análise do índice de nebulosidade para otimização do processo de agrupamentos de dados / Ernani Cláudio Borges. - 2012.

117 f.: il.

Orientador: Adriano Alves Pereira.

Coorientadora: Selma Terezinha Milagre.

Dissertação (mestrado) – Universidade Federal de Uberlândia, Programa de Pós-Graduação em Engenharia Elétrica.

Inclui bibliografia.

1. Engenharia elétrica - Teses. I. Pereira, Adriano Alves. II. Milagre, Selma Terezinha. III. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

CDU: 621.3

Dedico esta Dissertação:
aos meus Pais, João e Célia, por serem me porto seguro;
à Débora e meus filhos, Ian e Bianca, pelos meus momentos de ausência;
aos meus Irmãos, Erlon, Erlene e Érica, que mesmo longe compartilharam das dificuldades;
ao saudoso Matheus Oliveira Rodrigues.

Meus agradecimentos,

a Deus, cujo infinito amor e bondade concedeu-me a serenidade e perseverança para a realização de mais este sonho;

aos professores, Adriano e Selma, pelas orientações e pelos constantes estímulos, os quais, percebo que suas preocupações, chegam a se igualar às de meus pais, e isso muito me emociona. Deixo aqui o "Meu eterno agradecimento";

aos professores, Alexandre Cardoso, Edgard Afonso Lamounier Júnior, Luciano Vieira Lima, Antônio Cláudio Paschoarelli Vieira, Keiji Yamanaka, por compartilharem seus conhecimentos. E a querida Cinara Fagundes Paranhos Mattos, por ajudar na realização do mestrado;

aos colegas do BioLab - UFU, em especial: Bruno, Lucas, Branquinho, Marla, Reuder, Isabela, Mariana, Andrei, Alessandro, Maria Fernanda e Professora Ângela, pelos momentos de descontração, minimizando as tensões dos estudos;

aos amigos do MINTER que, tendo em vista todas as dificuldades, nos unimos, e juntos nos fortalecemos para a concretização deste trabalho. Assim, deixo aqui o meu forte abraço a todos;

ao Reitor, Roberto Gil, aos gestores do IFTM, aos companheiros do Campus Uberaba, pelo apoio, compreensão e liberação em momentos importantes de estudos;

aos Amigos do Campus Avançado Patrocínio por entenderem minha ausência e pela verdadeira amizade. Deixo um abraço especial ao amigo, Rogério Melo Nepomuceno, por assumir responsabilidades as quais não tenho como expressar minha gratidão;

a todos, que direta ou indiretamente, contribuíram na realização deste trabalho.

"Talvez não tenhamos conseguido fazer o melhor, mas lutamos para que o melhor fosse feito. Não somos o que deveríamos ser, não somos o que iremos ser ... mas Graças a Deus, não somos o que éramos."

Martin Luther King

RESUMO

BORGES, E.C. *Análise do índice de nebulosidade para otimização do processo de agrupamentos de dados*. Dissertação de Mestrado. Faculdade de Engenharia Elétrica da Universidade Federal de Uberlândia. Uberlândia, 2012.

A técnica de análise de agrupamento (*clustering analysis*) é uma ferramenta importante na pesquisa científica, podendo ser utilizada em diversas áreas do conhecimento tais como medicina, biologia e estatística. Agrupar dados é uma forma de refletir a estrutura interna dos dados e identificar classes presentes nesses agrupamentos, de modo que haja homogeneidade dentro das mesmas classes e heterogeneidade entre classes diferentes. Existem vários métodos de agrupamentos utilizados para encontrar o particionamento ótimo, dentre estes pode-se destacar: os métodos hierárquicos, métodos baseados em teorias dos grafos e métodos baseados em função objetivo. Neste trabalho foi utilizado o algoritmo baseado na função objetivo *Fuzzy C-Means* em conjunto com a técnica de reamostragem *bootstrap*. A ideia é variar o índice de nebulosidade para encontrar a melhor faixa de valores a ser utilizada para a classificação dos dados e consequentemente obtenção de melhores particionamentos. A qualidade da classificação é baseada em medidas de comparação tradicionais tais como Classificação Cruzada (Acc), F1, Hubert (Hub), Jaccard, Índice Randômico (Rand) e *Fowlkes and Mallows* (*Fowlkes*). As bases de dados utilizadas foram a *Iris*, *Wine* e três bases de dados artificiais. Os resultados obtidos demonstram que a melhor faixa de valor para o índice de nebulosidade está entre *1,04 e 1,2* para as medidas e bases de dados estudadas.

Palavras-Chave – Agrupamento de Dados, Índice de Nebulosidade, Fuzzy C-Means.

ABSTRACT

BORGES, E.C. *Analysis of cloudiness index for process optimization of data arrays*. Masters Dissertation. Faculty of Electrical Engineering. Federal University of Uberlândia. Uberlândia – Brazil, 2012.

The technique of clustering analysis is an important tool in scientific research, it can be used in various fields of knowledge such as medicine, biology and statistics. To group data in clusters is a way to reflect the internal data structure and identify classes present in this clusters so within the same class there is homogeneity and there is heterogeneity between different classes. There are three types of clustering methods used to find optimal partitioning: hierarchical methods, methods based on graph theory and methods based on objective function. In this study we used the objective function algorithm based on Fuzzy C-Means and also the bootstrap resampling technique. The idea is to vary the cloudiness index in order to find the best value to be used for sorting the databases: Iris, Wine and three other artificial databases, consequently obtaining better partitioning results. The quality of the partitioning is based on traditional measures of comparison such as Crusade Classification (Acc), F1, Hubert (Hub), Jaccard, Random index (Rand) and *Fowlkes and Mallows* (*Fowlkes*). The results obtained so far show that the best range for the cloudiness index is between *1.04* and *1.2* for the contents of measures adopted.

KEY WORDS - Clustering Analysis, Weighting Exponent, Index Cloudiness, Fuzzy C-Means

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

FCM	*Fuzzy* c-Means
NaN	Not a Number
Acc	Classificação Cruzada
Rand	Índice Randômico
F&M	Índice *Fowlkes and Mallows*
MDS	Muldimensional Scaling
Atrib	Atributo
Nro	Número

LISTA DE SÍMBOLOS

$\vec{X}$	base de dados
$\vec{x}$	dado
N	número de dados na base de dados
P	número de atributos que compõem o dado
C	número de grupos nebulosos
W	índice do atributo
K	índice do atributo
I	índice do grupo
R	índice do grupo
J	índice do dado
L	índice do dado
μ	grau de pertinência do dado dentro do grupo
μ_{ij}	grau de pertinência do dado j no grupo i
d_{ij}	distância do dado j ao centro do grupo i
T	t-norma
T_{minnp}	t-norma nil-potent mínimo
T_{prod}	t-norma produto
$\mathcal{X}^p$	espaço p-dimensional
$\vec{U}$	matriz de graus de pertinência nebulosos
$\vec{V}$ $\vec{U}$	matriz de centros dos grupos nebulosos
J_m	função objetivo do algoritmo *Fuzzy* c-*Means*
m	fator de fuzificação
t	número de iterações
$\vec{v}$	matriz que contém os centros nebulosos
ε	critério de parada para o algoritmo *Fuzzy* c-*Means*
p	é a dimensão do dado ou o espaço do dado
n_{11}	Verdadeiros positivos
n_{01}	Falsos positivos

n_{10}	Falsos negativos
n_{00}	Verdadeiros negativos
$\varsigma(i)$	Linha permutada *i*
ψ_{jl}	Valor de coincidência entre o dados *j* e *l*
Ψ	Matriz de coincidência
$\Pi(c)$	conjunto de todas as permutações dos *c* grupos

SUMÁRIO

CAPÍTULO 1

1 INTRODUÇÃO

Todos os dias as pessoas encontram uma grande quantidade de informações e as armazenam para posterior análise e gestão. Uma das ferramentas que podem ser utilizadas para estudo desses dados é a análise de agrupamentos (*clustering analysis*), que é uma das mais antigas técnicas em que não são feitas suposições com relação ao número de grupos ou à estrutura existente dentro do grupo. Os procedimentos exploratórios são frequentemente úteis para o entendimento da natureza complexa, existente nas relações multivariadas. Buscar nos dados uma estrutura de agrupamentos naturais é uma importante técnica exploratória, pois agrupamentos podem fornecer um meio informal para acesso à dimensionalidade, identificando tendências e sugerindo hipóteses relativas às semelhanças (JOHNSON, 1992).

A partir das técnicas de agrupamento, diversos estudos a aplicam como uma ferramenta para análise, tendo sido utilizada nas áreas de processamento de imagens, biologia, reconhecimento de dados, mineração de dados, sensoriamento remoto, bioinformática, dentre outras (JAIN, MURTY, FLYNN, 1999).

Existem três métodos de agrupamentos que são utilizados para encontrar o particionamento ótimo: métodos "hierárquicos", métodos baseados em "teoria dos grafos" e métodos baseados em "função objetivo". Os métodos baseados em função objetivo são muito utilizados (SARKAR, LEONG, 2001) e dentre eles pode-se destacar a Classificação Nebulosa (*Fuzzy*).

Vários algoritmos de classificação têm sido propostos (JAIN, DUBES, 1998) para descobrir automaticamente o particionamento *natural* da base de dados. Porém, os resultados de particionamento que os algoritmos geram, não se sabe se correspondem ao número correto de grupos.

Assim, independentemente dos algoritmos de classificação propostos, os resultados devem ser validados de forma quantitativa e objetiva. Uma das maneiras existentes para a validação é tomar como base as estabilidade da solução encontrada (ROTH, LANGE, 2002). Essas validações podem ser de dois tipos de critérios: um "externo", onde compara-se os resultados do classificador com informações que estão fora da base de dados; e o outro critério é "interno" o qual usa a própria base de dados, isso implica dizer que deve haver a reamostragem dos dados.

Dessa forma, a reamostragem pode ser utilizada para escolha do particionamento mais consistente presente na base de dados. A ideia é realizar a comparação entre a base de dados completa e sub-amostras dessa base de dados. Espera-se que, quando o número de grupos estiver correto, as sub-amostras e a base de dados original tenham a mesma estrutura de grupos. Para um número incorreto de grupos o resultado do agrupamento deve ser instável (minimizando os valores das medidas utilizadas) (ROTH, LANGE, 2002; BORGELT, 2006).

Porém, a estabilidade em uma Classificação *Fuzzy* não pode ser definida simplesmente pelo método de reamostragem, pois, nessa classificação existe um parâmetro chamado "expoente de ponderação *m*" (também conhecido por índice de nebulosidade ou índice de fuzificação), que é empírico e seu valor influencia substancialmente os resultados da classificação nebulosa (XU, WUNSCH II, 2005).

Neste contexto, este trabalho realizou um estudo para buscar a melhor faixa de valores de *m*, ou seja, que leve aos agrupamentos mais estáveis nas bases de dados e medidas analisadas. A qualidade de comparação é baseada em medidas de comparação tradicionais tais como Classificação Cruzada (Acc), F1, Hubert (Hub), Jaccard, Índice Randômico (Rand) e *Fowlkes and Mallows* (*Fowlkes*). Foram utilizadas as bases de dados *Iris*, *Wine*, *Brady1* e três bases de dados artificiais.

CAPÍTULO 2

2 AGRUPAMENTO DE DADOS

A análise de agrupamentos (*clustering analisys*) é uma das mais antigas técnicas em que não são feitas suposições com relação ao número de grupos ou à estrutura existente dentro do grupo. Buscar nos dados uma estrutura de *agrupamentos naturais* é uma importante técnica exploratória, pois agrupamentos podem fornecer um meio informal para acesso à dimensionalidade, identificando tendências e sugerindo hipóteses relativas às semelhanças. É uma ferramenta importante e tendo diversos estudos que aplicam esta técnica, como por exemplo processamento de imagens, biologia, reconhecimento de dados, mineração de dados, sensoriamento remoto, bioinformática, etc (JOHNSON, 1992).

A análise de agrupamento é um conjunto de padrões, geralmente representados por um vetor de medições multidimensional, com base na similaridade dos grupos (JAIN; MURTY; FLYNN, 1999).

Basicamente, os sistemas de classificação são supervisionados ou não supervisionados (XU, 2005). Em classificação supervisionada tem-se uma coleção de dados rotulados (pré-classificados), sendo que o problema é rotular um novo dado encontrado e ainda não rotulado. Normalmente, os dados marcados (padrão) são usados para aprender as descrições de classes que por sua vez são utilizados para rotular um novo padrão. Já a classificação não supervisionada o problema é agrupar um dado conjunto de dados não rotulados dentro de um agrupamento significativo. Neste sentido, os rótulos são associados também com grupos, nesta categoria, rótulos são dirigidos a dados, isto é, eles são obtidos exclusivamente a partir dos dados. (JAIN, MURTY, FLYNN, 1999).

Para encontrar um particionamento ótimo os métodos usualmente utilizados são: métodos "hierárquicos", métodos baseados em "teoria dos grafos" e métodos baseados em "função objetivo". Os métodos baseados em função objetivo são muito utilizados (SARKAR; LEONG, 2001) e dentre eles pode-se destacar a Classificação Nebulosa (*Fuzzy*).

Essencialmente, um algoritmo de agrupamento executa a tarefa de particionar uma base de dados em um número de grupos homogêneos, respeitando um grau de similaridade. A principal diferença entre agrupamento tradicional (*hard* ou *crisp*) e de lógica nebulosa (*Fuzzy*) é que no tradicional uma característica pertence a um grupo e em lógica *Fuzzy* podem pertencer a vários grupos com graus diferentes de distância (XIE; BENI, 1991; HÖPPNER *et al*, 1999).

Diferentes algoritmos de agrupamento foram propostos para diferentes aplicações (JAIN; DUBES, 1988), e independente do algoritmo, os resultados devem ser validados de forma quantitativa e objetiva. Uma das maneiras existentes para a validação é tomar como base as estabilidades da solução encontrada (ROTH *et al*, 2002).

A principal preocupação em um processo de agrupamento é revelar a organização dos padrões dentro de grupos sensíveis, os quais nos permitem descobrir similaridades e diferenças, assim como derivar inferências úteis sobre eles (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a).

O critério de validação utiliza exclusivamente a própria base de dados, isso implica dizer que deve haver a reamostragem dos dados.

Assim, a reamostragem pode ser utilizada para escolha do particionamento mais consistente presente na base de dados. A ideia é realizar a comparação entre a base de dados completa e sub-amostras dessa base de dados. Espera-se que, quando o número de grupos estiver correto, as sub-amostras e a base de dados original tenham a mesma estrutura de grupos. Para um número incorreto de grupos o resultado do agrupamento deve ser instável minimizando os valores das medidas utilizadas (ROTH et al, 2002; BORGELT, 2006).

As pessoas sempre tentam buscar recursos que possam auxiliar na tomada de decisão. Assim, a análise de agrupamento de dados tende a agrupar dados com base na similaridade ou dissimilaridade (distância) de acordo com determinadas normas ou regras (XU, 2005).

A análise de agrupamento (Figura 2.1) consiste basicamente em:

- "Representação padrão" extrai (seleciona) características distintas de um conjunto de objetos. São utilizadas algumas regras para gerar os novos dados a partir dos dados originais. Geralmente, as características devem conter padrões distintos pertencentes a diferentes grupos (XU, 2005);

- "Projeto de agrupamento" é a combinação da seleção de uma medida de proximidade, a qual afeta diretamente a formação dos grupos e o desenvolvimento das rotinas de critério de semelhanças (agrupamentos), pois, quase todos os algoritmos de agrupamentos são explícita ou implicitamente ligados a alguma definição de medida de proximidade e não há um algoritmo universal para todos os problemas (XU, 2005);

- "Validação do agrupamento" é responsável por avaliar a saída do procedimento de agrupamento, pois, a identificação de parâmetros ou a ordem de apresentação dos padrões de entrada podem afetar os resultados finais. Portanto, normas de avaliação e critérios são importantes para fornecer aos usuários resultados com alto grau de confiabilidade (XU, 2005);

- "Interpretação dos resultados" deverá ser capaz de fornecer aos usuários, conhecimentos significativos a partir dos dados originais, possibilitando a tomada de decisões e solução de problemas (XU, 2005).

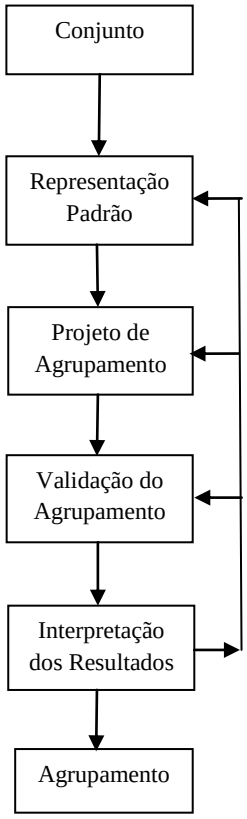


Figura 2.1 - Fases do agrupamento de dados (modificada de XU, 2005)

Vale ressaltar a importância da validação do procedimento de agrupamentos, pois, dados que não possuem grupos não devem ser processados por um algoritmo de agrupamento. Assim, os estudos chamados de *tendência de classe* (*cluster tendency*) analisam a entrada verificando a existência de mérito para análise de agrupamento. A "análise de validade de grupos" (*cluster validity analysis*) avalia a saída do procedimento determinando se a estrutura do agrupamento é válida, para tal, a estrutura não pode ter ocorrido por acaso ou como um artefato do algoritmo. (JAIN; MURTY; FLINN, 1999).

Segundo Roth (et al, 2002), a validação dos resultados de forma quantitativa e objetiva pode ser com dois critérios: externo e/ou interno, sendo que o critério interno usa exclusivamente o próprio conjunto de dados, e o critério externo a solução do classificador é comparada com informações que não estão na base de dados.

2.1 Definições e Notações

A Tabela 2.1 apresenta definições e notações que serão apresentados no decorrer deste trabalho

Tabela 2.1 - Definições e Notações (JAIN, MURTY, FLYNN, 1999)

Cluster	**Agrupamento**
Dado	vetor de características ou vetor de atributos; vetor que possui *p* medidas (ou atributos): $\vec{x} = (x_1, ..., x_p)$;
Hard	Técnica de agrupamento fixa
Fuzzy	Técnica de agrupamento nebulosa
Medida de distância	Métrica utilizada para quantificar o padrão de semelhança.
Atributo	ou características são os componentes individuais do dado x;
Dimensionalidade	*p* é a dimensão do dado ou o espaço do dado;
Base de dados	são todos os dados: **X** = $(\vec{x}_1, ..., \vec{x}_n)$. O *j-ésimo* dado em **X** é denotado por $\vec{x} = (x_{j1}, ..., x_{jp})$, ou seja, o conjunto total de dados **X**, com *p* atributos e *n* dados;
Grupo	é um conjunto de dados similares entre si e diferentes dos outros dados presentes na base de dado.

A Tabela 2.2 demonstra, de forma simbólica, a composição do conjunto **X** e sua totalidade de dados, contendo *p* atributos e *n* dados.

Tabela 2.2 - Demonstração do conjunto de dados **X**

	Atributo 1	***…***	***Atributo w***	***…***	***Atributo p***
Dado 1:	x_{11}	…	x_{1w}	…	x_{1p}
Dado 2:	x_{21}	…	x_{2w}	…	x_{2p}
…	…	…	…	…	…
Dado j:	x_{j1}	…	x_{jw}	…	x_{2j}
…	…	…	…	…	…
Dado n:	x_{n1}	…	x_{nw}	…	x_{np}

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1w} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2w} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jw} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nw} & \cdots & x_{np} \end{pmatrix} \quad (1)$$

Em outras palavras, vamos supor que a matriz $\vec{X}$ fosse composta por um diagnóstico médico, as linhas (dado) seriam os pacientes e as colunas (Atributo) seriam os sintomas (características como pressão, temperatura, etc).

2.2 *Lógica Fuzzy*

Também conhecida como "Lógica Nebulosa" foi introduzida em 1965 pelo matemático, Lofti Asker Zadeh, por meio da publicação de um trabalho sobre Conjuntos *Fuzzy*, baseado na lógica multinível, o qual mostra o tratamento dos aspectos imprecisos e incertos. A lógica tradicional (*hard* ou *crisp*) trata os valores 0 e 1 (falso ou verdadeiro) não podendo ser verdadeira e falsa ao mesmo tempo. Na lógica *Fuzzy* é possível encontrar os valores "entre 0 e 1" podendo ser quase falso como também quase verdadeiro, permitindo assim, explorar a tolerância à imprecisão, à incerteza e à veracidade parcial para alcançar tratabilidade, robustez (ZADEH, 1994).

A teoria do conjunto nebuloso diz que, dado um determinado elemento que pertence a um domínio, é verificado o grau de pertinência (também chamada de função característica) do elemento em relação ao conjunto. O grau de pertinência é a referência para verificar o quanto é possível esse elemento poder pertencer ao conjunto. Um conjunto nebuloso **A** na base de dados **X** é caracterizado por uma função de pertinência $\mu_A(\vec{x})$ a qual associa cada ponto em **X** a um número real no intervalo [0,1], com o valor de $\mu_A(\vec{x})$ representando o grau de pertinência de x em **A**, ou seja, **A** é um conjunto de pares ordenados do elemento genérico *x*, dado pela Equação 2.1:

$$A = \left\{ (\vec{x}, \mu_A(x)) \mid \vec{x} \in X \right\} \tag{2.1}$$

se $\mu_A(\vec{x})$ = 1 tem-se pertinência total ao conjunto nebuloso **A**, e $\mu_A(\vec{x})$ = 0, não se tem pertinência ao conjunto nebuloso **A**. Um valor próximo de zero indica "baixo" grau de pertinência e um valor próximo de 1, indica "alto" grau de pertinência.

A Figura 2.2 exemplifica agrupamentos nebulosos e não nebulosos sendo os retângulos H1 e H2 agrupamentos *crisp* (*hard*) e as elipses F1 e F2 a saída do algoritmo nebuloso.

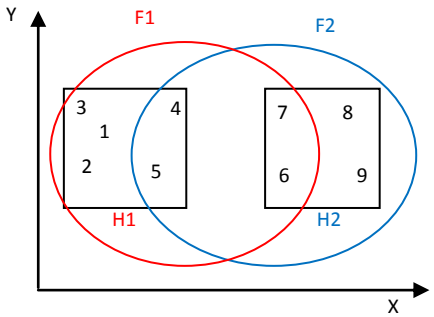


Figura 2.2. Agrupamento nebuloso (modificado de JAIN, MURTY, FLYNN, 1999)

2.3 *Operadores de Intersecção e União*

Os conjuntos nebulosos, de forma semelhante aos conjuntos "*crisp*" necessitam executar operações, tais como: intersecção, união, negação dentre outras. Assim, segue algumas definições para os conjuntos nebulosos (ZADEH, 1965):

Seja os conjuntos A, B e C definidos em **X**:

a) É vazio *se e somente se* sua função de pertinência for idêntica a zero em **X**;

b) A e B são iguais, *se e somente se* $\mu_A(\vec{x}) = \mu_B(\vec{x})$ para todo x em **X**;

c) Complemento de A (denotado por A') é definido por: $\mu_{A'}(\vec{x}) = 1 - \mu_A(\vec{x})$;

d) $A \subset B$, *se e somente se* $\mu_A(\vec{x}) \leq \mu_B(\vec{x})$: $A \subset B \Leftrightarrow \mu_A(\vec{x}) \leq \mu_B(\vec{x})$;

e) União de A e B, tem-se: $C = A \cup B$ ... assim:

$$\mu_C(\vec{x}) = Max[\mu_A(\vec{x}), \mu_B(\vec{x})], \quad \vec{x} \in \mathbf{X}$$

f) Intersecção de A e B, tem-se: $C = A \cap B$ ... assim:

$$\mu_C(\vec{x}) = Min[\mu_A(\vec{x}), \mu_B(\vec{x})], \quad \vec{x} \in \mathbf{X}$$

g) Produto algébrico de A e B, tem-se: AB ... assim:

$$\mu_{AB}(\vec{x}) = \mu_A(\vec{x})\mu_B(\vec{x})$$

h) Soma algébrica de A e B, tem-se: A + B ... assim:

$$\mu_{A+B}(\vec{x}) = \mu_A(\vec{x}) + \mu_B(\vec{x})$$

i) Diferença absoluta de A e B, tem-se: |A – B| ... assim:

$$\mu_{|A-B|}(\vec{x}) = |\mu_A(\vec{x}) - \mu_B(\vec{x})|$$

2.4 Normas Triangulares

As normas triangulares foram introduzidas para modelar distância no espaço métrico probabilístico (MENGER, 1942). Quando o conjunto é *crisp*, as operações são de intersecção e união, já em conjuntos nebulosos a união é representada pelo operador de *t-conormas* ($\perp$) e a intersecção pelo operador de *t-normas* (T).

Essas normas são definidas como: T*:[0,1] x [0,1]* $\rightarrow$ *[0,1]*, operações associadas às norma triangular são definidas para $\forall a,b \in [0,1]$ sendo muito utilizadas na prática como implicações em aplicações de controle nebuloso (SANDRI; CORREA, 1999).

a) Comutatividade: $\mathsf{T}(a,b) = \mathsf{T}(b,a)$;

- observa-se que T independe da ordem dos conjuntos nebulosos

b) Associatividade: $\mathsf{T}(a,\mathsf{T}(b,c)) = \mathsf{T}(\mathsf{T}(a,b),c)$;

- generaliza o operador T para qualquer número de conjuntos nebulosos

c) Monotonicidade: $\mathsf{T}(a,b) \leq \mathsf{T}(c,w)$ se $a \leq c$ e $b \leq w$;

- implica que um decréscimo nos valores de pertinência de dois conjuntos nebulosos não pode produzir um aumento no valor de pertinência após a aplicação do operador T.

d) elemento neutro = 1: $\mathsf{T}(a,1) = a$;

elemento neutro = 0: $\perp(a,0) = a$;

e) $\mathsf{T}(0,0) = 0$.

Os elementos neutros e $\mathsf{T}(0,0)=0$, indicam que as *t-normas* podem ser generalizadas para conjuntos clássicos.

A Tabela 2.3 apresenta as *t-normas* e *t-conormas* mais utilizadas em conjuntos nebulosos.

Tabela 2.3 - Principais T-normas

t-norma	**Nome**
min(a,b)	Zadeh
a . b	Produto
Min(a,b), se a+b ≥ 1, 0 senão	*nil-potent* mínimo
max(a+b – 1,0)	Lukasiewicz

Obs.: Neste trabalho utilizou-se as t-normas *nil-potent* mínimo ($T_{minnp}(a,b)$) e produto ($T_{prod}(a,b)$)

2.5 Algoritmo Fuzzy C-Means (FCM)

É um típico algoritmo de análise de grupos e tem sido amplamente utilizado na área científica (ZHANG, CHEN, 2003). Suas principais características são baixa complexidade computacional e facilidade na implementação e sua técnica de análise de grupos é baseada em *função objetivo* a qual atribui a cada partição um valor de que deve ser otimizado, assim obtendo a melhor avaliação (NUOVO, et al, 2006; HÖPPNER, 1999). É o equivalente nebuloso do algoritmo *crisp Fuzzy K-Means* (XIE; BENI, 1991).

O processo de cálculo do algoritmo FCM (Equação 2.2) é iterativo. Assim, o propósito é minimizar o índice de desempenho da pseudopartição nebulosa J_m (ou *função objetivo*) que mede a distância entre os centros dos grupos e os elementos dentro dos grupos, desta forma, quanto menor seu valor, mais otimizada estará a pseudo-partição ou partição nebulosa U.

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m (d_{ij})^2 \qquad (2.2)$$

Onde:

$\vec{U}$ - matriz de graus de pertinência nebulosos.

$\vec{V}$ - conjunto de vetores que representa os *c* centros dos grupos, ou seja,

$$V = \{v_1, v_2, ..., v_c\}$$

c - número de grupos, sendo inteiro, positivo e maior que 1 e menor que *n* (número de dados).

μ_{ij} - é a pertinência do dado *j* no grupo *i*.

d_{ij} - é a distância do dado *j* ao centro do grupo *i*.

m - é o expoente de ponderação (índice de nebulosidade).

O valor de *c* deve ser inteiro, maior que 1 e menor que *n* dados, pois se for igual a 1, todos os dados iriam pertencer ao mesmo grupo e cada centro coincidiria exatamente com um dado e a pertinência seria igual a 1 nesse grupo e zero nos demais tornando-se um particionamento *crisp* (rígido). Essa regra satisfaz a seguinte equação (Equação 2.3) onde a soma de todos os graus de pertinência de um determinado dado deve ser igual a 1 no total de agrupamentos.

$$\sum_{i=1}^{c} \mu_i(\vec{x}_j) = 1, \forall j \in \{1,2,...,n\} \tag{2.3}$$

As Equações 2.4 e 2.5 representam o cálculo da distância, a qual, normalmente utiliza-se como métrica a distância Euclidiana:

$$d_{ij} = d(\vec{x}_j - \vec{v}_i) \tag{2.4}$$

$$d_{ij}^2 = \| \vec{x}_j - \vec{v}_i \|^2 = \sum_{k=1}^{p} (x_{jk} - \upsilon_{ik})^2 \tag{2.5}$$

Assim, o problema do agrupamento nebuloso é encontrar os centros dos agrupamentos associados que represente a estrutura dos dados da melhor forma possível. Sendo esse objetivo alcançado quando as associações dos dados são fortes dentro do agrupamento e fracas entre os agrupamentos. Desta forma, define-se um índice de desempenho baseados nos centros dos agrupamentos. Tem-se uma dada pseudo-partição $U = \{\mu_{1j}, \mu_{2j}, ..., \mu_{cj}\}$ e o centro de agrupamento *c*, pode-se calcular os centros associados a cada partição como demonstra a Equação 2.6:

$$\vec{v}_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m (\vec{x}_j)}{\sum_{j=1}^{n} (\mu_{ij})^m} \tag{2.6}$$

Onde μ_{ij} é o grau de inclusão do dado *j* no grupo *i*, para $\forall i \in \{1,2, ..., c\}$ e $m \in (1, \infty)$ é um número real que representa o fator de nebulosidade e define a faixa de nebulosidade existente entre um grupo e outro. Em outras palavras, quando o valor de m $\rightarrow \infty$, as pertinências tendem a 1/*c*, assim, os pontos têm o mesmo grau de inclusão em todos os grupos desta forma o grau de pertinência fica mais nebuloso, caso o valor de $m \rightarrow 1$, levaria a uma matriz totalmente rígida (*crisp*) e o algoritmo *Fuzzy c-Means* converge para o *Hard k-Means*.

Basicamente os passos para execução do algoritmo *Fuzzy C-Means* para uma base de dados são: definir (aleatoriamente) o índice de nebulosidade *m* (maior ou igual a 1, tendendo ao infinito), estabelecer o número de agrupamentos *c* (maior que 1 e menor que a quantidade dados), definir o critério de parada ($\varepsilon > 0$), o número máximo de iterações *t (t = 0,1,2,3,...)* inicializando-se a matriz de partições nebulosas $U^{(0)} = [\mu_{ij}]$, calcular vetor de protótipo (v_i conforme equação 2.6) ou seja, cálculo dos centros dos grupos $V^{(t)}$), atualizar os graus de pertinências (μ_{ij}) conforme a Equação 2.7:

$$\mu_{ij} = \frac{1}{\sum_{r=1}^{c} \left(\frac{d^2(\vec{x}_j, \vec{v}_i)}{d^2(\vec{x}_j, \vec{v}_r)} \right)^{\left(\frac{1}{m-1}\right)}} \tag{2.7}$$

Os agrupamentos obtidos pelo algoritmo FCM devem ser validados, pois como os demais algoritmos de agrupamento ele produz um modelo de particionamentos para uma base de dados, quer existam ou não. Essa avaliação pode ser feita verificando se o modelo de agrupamento obtido é o que mais se adéqua ao conjunto de dados ou avaliando-se a qualidade do agrupamento.

Em geral, validar soluções de *cluster* significa avaliar os resultados da análise de *cluster* de forma quantitativa e objetiva. Tal avaliação pode ser baseada em dois tipos de critérios: (i) os critérios externos: uma solução de *cluster* corresponde a uma informação a priori, ou seja, a informação externa que não está contida no conjunto de dados. (ii) critérios internos: a medida de qualidade é exclusivamente baseada nos próprios dados. Esse critério de validação utiliza exclusivamente a própria base de dados, isso implica dizer que deve haver a reamostragem dos dados (ROTH et al, 2002).

Dessa forma, a reamostragem pode ser utilizada para escolha do particionamento mais consistente presente na base de dados. A ideia é realizar a comparação entre a base de dados completa e sub-amostras dessa base de dados. Espera-se que, quando o número de grupos estiver correto, as sub-amostras e a base de dados original tenham a mesma estrutura de

grupos. Para um número incorreto de grupos o resultado do agrupamento deve ser instável (minimizando os valores das medidas utilizadas) (ROTH et al, 2002; BORGELT, 2006).

Porém, a estabilidade em uma Classificação *Fuzzy* não pode ser definida simplesmente pelo método de reamostragem, pois, nessa classificação existe um parâmetro chamado "expoente de ponderação *m*" (também chamado de índice de nebulosidade ou índice de fuzificação), que é empírico e seu valor influencia substancialmente os resultados da classificação nebulosa (PAL; BEZDEK, 1995). A precisão pela qual o algoritmo classifica os dados depende fortemente do valor de *m*.

2.6 Reamostragem Booststrapping

A reamostragem com reposição *bootstrapping* (EFRON; TIBSHIRANI, 1993) pode ser usada para estimar índices relativos. A comparação é feita entre o agrupamento resultante para uma dada amostra da base de dados e um agrupamento de referência. Assim, os métodos baseados em reamostragem são utilizados para verificar onde a base de dados possui uma estrutura de agrupamentos ou onde o resultado é somente um artefato do algoritmo e também para selecionar-se o modelo do número de grupos (BORGELT, 2005; LAW, JAIN, 2003).

2.7 Validade do agrupamento

Vários algoritmos de agrupamento baseiam-se em algumas suposições para definir o particionamento de uma base de dados, o que pode levar a diferentes resultados dependendo das características do conjunto de dados, como por exemplo geometria e densidade de distribuição dos grupos, e dos valores dos parâmetros de entrada (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a).

Considerando que algoritmos de agrupamentos produzem um modelo de particionamento para uma base de dados, torna-se necessária a validação desses agrupamentos. O procedimento de avaliação que pode fornecer uma resposta quantitativa é denominado validade do agrupamento (*cluster validity*), o qual é o objetivo de muitos esforços de pesquisadores (LAW; JAIN, 2003; PAL; BEZDEK, 1995; XIE; BENI, 1991; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a; HALKIDI, BATISTAKIS; VAZIRGIANNIS, 2002b).

A validação do agrupamento é responsável por avaliar a saída do procedimento de agrupamento, pois a identificação de parâmetros ou a ordem de apresentação dos padrões de entrada podem afetar os resultados finais. Portanto, normas de avaliação e critérios são importantes para fornecer aos usuários resultados com alto grau de confiabilidade (XU, 2005).

Essas validações podem ser de dois tipos de critérios: um "externo", onde compara-se os resultados do classificador com informações que estão fora da base de dados; e o outro critério é "interno" o qual usa a própria base de dados.

No ANEXO A, são apresentados os cálculos necessário para comparação de matrizes de partição $U^{(1)}$ e $U^{(2)}$, tabelas de contingência para comparar duas linhas de duas matrizes de partição *crisp* e aplicação das normas triangulares.

CAPÍTULO 3

3 MATERIAIS E MÉTODOS

Os testes foram desenvolvidos em um notebook Dell, modelo Vostro 3450, com processador Intel® core™ i5-2410M CPU@2.30 GHz, 6 GB de Memória RAM, HD 500GB, com sistema operacional Windows 7 - 64 bits e software MATLAB®.

As análises foram feitas em seis bases de dados: Artificial 1, Artificial 2, Artificial 3, *Iris*, *Wine* e *Brady 1*. As bases de dados *Iris*, *Wine* são bases de dados reais e foram obtidas no repositório *UCI Machine Learning Repository* (BALKE; MERZ, 1998). A base de dados *Brady 1* compõe um banco de dados genotípico de bactérias simbióticas fixadoras de nitrogênio, coletivamente chamada de rizóbios (HUNGRIA; VARGAS; ARAÚJO, 1997).

O método teve início com a normalização da base de dados (dados originais) para média zero e desvio padrão unitário, para remover os efeitos de escala. Na execução do algoritmo FCM foi utilizado um número de iterações de t = 100; índice nebulosidade m = [1,01 1,02 1,03 1,04 1,05 1,06 1,07 1,1 1,2 1,5 1,8 2,0 2,3 2,5 3,0 3,5 4,0 4,5]; critério de parada $\varepsilon = \leq 0{,}00001$ e o número de grupos c = 2, 3, ..., 8.

O número de grupos variou de 2 a 8, para permitir uma melhor análise em torno do número de agrupamentos existentes nas bases de dados analisadas. Foram utilizadas 100 amostras e o método de reamostragem com reposição (*bootstrapping*) foi aplicado para geração da sub-amostra de dados contendo 90% dos dados da referência (base de dados original). A Tabela 3.1 demonstra as características das bases de dados

Tabela 3.1 – Características das bases de dados

Conjunto de Dados	**Número de Dados**	**Número de Atributos**	**Número de grupos**
Artificial 1	400	2	4
Artificial 2	350	2	7
Artificial 3	1000	3	5
Iris	150	4	3
Wine	178	13*	3
Brady 1	119	49**	5 ou 6

* neste trabalho só foram utilizados os atributos 7, 10 e 13.

** foram utilizados os atributos de 1 a 20

3.1 Bases de dados

3.1.1 Base de dados Artificial 1

Formada por um conjunto bidimensional artificial, contem 400 dados e 2 atributos divididos em 4 grupos. O primeiro grupos foi gerado com 100 dados e com 2 atributos, se iniciou com a função *rand* (MATHWORK, 1998). Demais dados, foram gerados acrescentando 1,2 ao valor do atributo 1 e 2,6 ao valor do atributo 2. Importante observar que os grupos estão linearmente separados, conforme demonstrado na figura 3.1.

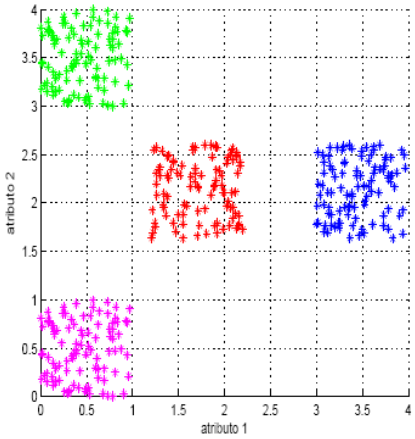


Figura 3.1 - Base de dados Artificial 1 (MILAGRE, 2008)

3.1.2 Base de dados Artificial 2

De forma similar a base de dados Artificial 1, esta base é formada por um conjunto bidimensional artificial, possui 350 dados, 7 grupos e 2 atributos. Seus atributos foram gerados por variação controlada de incrementos. Demais dados, foram gerados acrescentando 2,2 ao valor do atributo 1, e 1,2 ao valor do atributo 2. Observa-se na figura 3.2 que os dados formam grupos completamente separados.

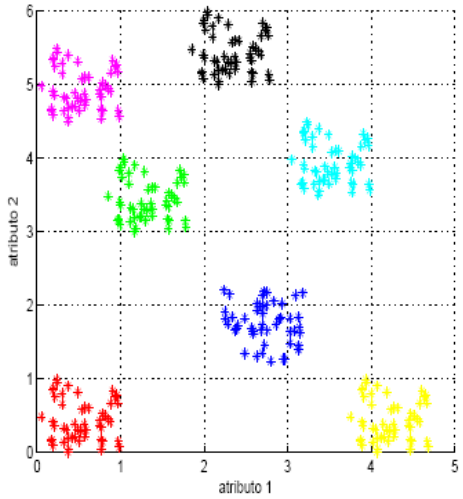


Figura 3.2 - Base de dados Artificial 2 (MILAGRE, 2008)

3.1.3 Base de dados Artificial 3

Gerada aleatoriamente por cinco misturas de distribuições gaussianas, tendo 1000 dados, 5 grupos e 3 atributos (TALAVERA, 2007). Observa-se que os grupos estão muito próximo e que alguns dados estão sobrepondo os dados de outros grupos. Na figura 3.3 fica visível esta sobreposição.

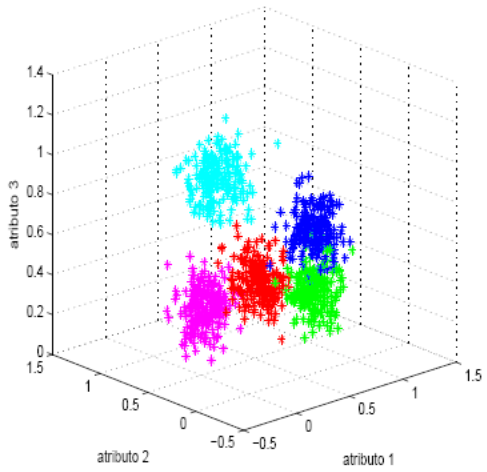


Figura 3.3 - Base de dados Artificial 3 (MILAGRE, 2008)

3.1.4 Base de dados Iris

Base de dados foi obtida no repositório *UCI Machine Learning Repository* (BALKE; MERZ, 1998). Seus dados representam a largura da sépala, comprimento da sépala, largura da pétala e comprimento da pétala da planta *Iris*, muito conhecida como "Flor de Lis". A base de dados é composta por 150 dados, 4 atributos cada, distribuídos em três grupos com 50 dados em cada grupo. Na Figura 3.4, percebe-se que um grupo está separado e os outros dois estão tão próximos que alguns dados estão sobrepostos.

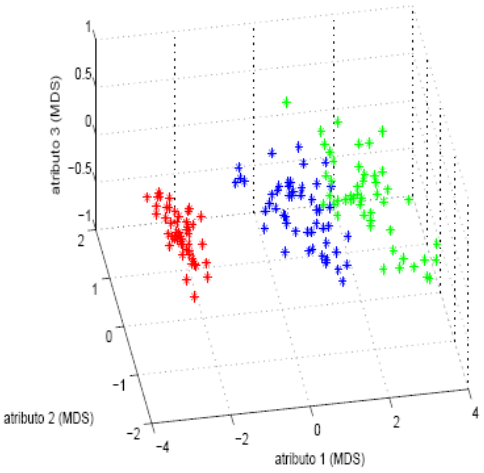


Figura 3.4 - Base de dados *Iris* (MILAGRE, 2008)

3.1.5 Base de dados Wine

Esta base também é real e foi obtida juntamente com a base de dados *Iris* do *UCI Machine Learning Repository* (BALKE, MERZ, 1998). Seus dados representam uma análise química de três classes de vinhos italianos, de uma mesma região, mas derivados de três diferentes cultivos (MILAGRE, 2008). Composta por 178 dados com 13 atributos cada. Para este trabalho utilizou-se os atributos 7, 10 e 13 por possuírem grupos mais representativos (TIMM; DÖRING, KRUSE, 2004; BORGELT, 2006). Na figura 3.5 observa-se que os três grupos estão muito próximo e os dados de um grupo está sobrepondo os outros dois.

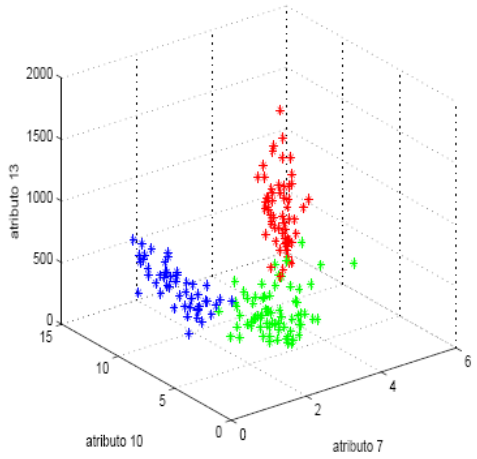


Figura 3.5 - Base de dados *Wine* (MILAGRE, 2008)

3.1.6 Base de Dados Bradyrhizobium

O nitrogênio é um elemento importante na síntese de proteínas, e tem-se uma demanda muito grande na agricultura. Esse nitrogênio pode ser absorvido diretamente do solo ou ser fornecido pela fixação biológica do nitrogênio, realizada por bactérias da família *Rhizobiacea*, tendo a relação simbiótica mais importante com bactérias pertencentes à espécie *Bradyrhizobium* (BOHRER, HUNGRIA, 1997).

A fixação biológica de nitrogênio é o sistema natural de transformação de N_2 atmosférico por bactérias que vivem no solo em formas assimiláveis pelas plantas. Quando associados a raízes, formam nódulos, onde ocorre a conversão do nitrogênio atmosférico em compostos nitrogenados, que são utilizados pela planta (MILAGRE, 2003).

Esse banco de dados foi gerado utilizando-se as regiões ribossomais 16S rRNA e o espaço intergênico 16S-23S (IGS) e nove enzimas de restrição, que cortam o DNA produzindo fragmentos de diferentes tamanhos tornando possível a comparação entre diferentes indivíduos, neste caso bactérias de solo (GERMANO et al., 2006).

Os dados utilizados neste trabalho, referentes a essa base de dados, foram obtidos de (LLERENA, 2008), ou seja são resultantes da aplicação da técnica MDS sobre os dados de

(MILAGRE, 2003). Será utilizado uma parte do banco de dados completo, referentes à região ribossomal 16S.

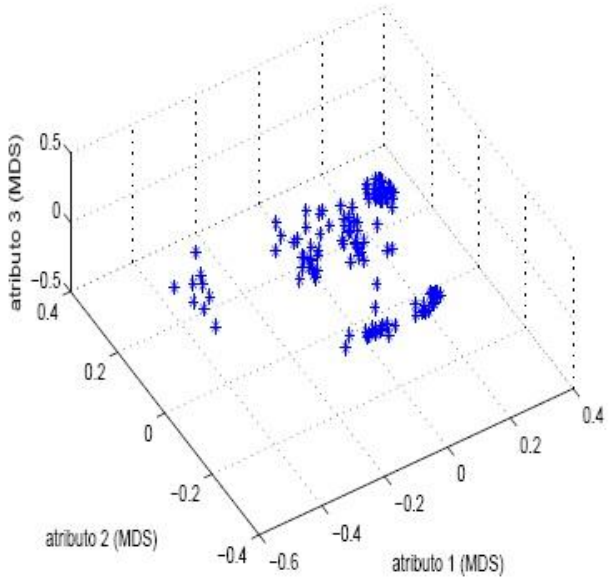


Figura 3.6 - Base de dados *Brady* 1 (MILAGRE, 2008)

A base de dados *Brady 1*, corresponde à região ribossomal 16S rRNA digerida com as enzimas de restrição Cfo I, Dde I, Msp I. Portanto essa base de dados contém dados referentes às três enzimas. Sendo 119 dados com 49 atributos cada.

O objetivo da utilização desta base de dados é a verificação e confirmação dos resultados.

3.2 Método

Neste trabalho foram feitas simulações e análises com seis bases de dados: Artificial 1, Artificial 2, Artificial 3, *Iris*, *Wine e Brady 1*.

O método teve início com a normalização da base de dados (original) para média zero e desvio padrão unitário, para remover os efeitos de escala. Essa normalização é feita toda vez que uma nova matriz de partição U é gerada. A Equação 3.1 mostra a restrição que deve ser satisfeita.

$$\sum_{i=1}^{c} \mu_i\left(\vec{x}_j\right) = 1, \forall j \in \{1,2,...,n\} \tag{3.1}$$

Na execução das simulações, utilizou-se do algoritmo *Fuzzy C-Means* (FCM) para realizar o agrupamento dos dados gerando a primeira matriz de comparação $\mathbf{U}^{(1)}$ ($\mathbf{U} = (\mu_{ij})$), a Tabela 3.2 apresenta os valores para os parâmetros utilizados:

Tabela 3.2 - Dados utilizados no algoritmo FCM

número de iterações	*t = 100*
Índice de nebulosidade	*m* = [1,01 1,02 1,03 1,04 1,05 1,06 1,07 1,1 1,2 1,5 1,8 2,0 2,3 2,5 3,0 3,5 4,0 4,5]
critério de parada	$\varepsilon \leq 0{,}00001$
Número de grupos	*c = 2, 3, 4, 5, 6, 7 e 8*

O método de reamostragem com reposição *bootstrapping* foi aplicado para geração de sub-amostra de dados contendo 90% dos dados da referência (base de dados original). Foram utilizadas 100 sub-amostras.

Após a execução do algoritmo *Fuzzy c-Means* e de posse das matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$, calcula-se as medidas de comparação binária tais como Classificação cruzada (Acc), F1, Hubert (Hub), Jaccard, Índice Randômico e *Fowlkes and Mallows* podem ser utilizadas para validação (MILAGRE, 2008).

Essas medidas comparam duas matrizes de partição e para esses cálculos aplicou-se a *t-norma produto* e para comparação entre matrizes de coincidência aplicou-se a combinação das *t-normas nil-potent* mínimo e produto.

O valor final para cada medida é obtido aplicando-se a média entre os resultados das 100 sub-amostras. O valor máximo obtido para cada medida, entre todos os particionamentos (*c* = 2, 3, ..., 8), indica o número de grupos definido pela medida como sendo o mais correto para a base de dados em análise.

A Figura 3.7 apresenta o fluxograma para o cálculo do número de grupos em bases de dados, utilizando reamostragem *bootstraping* e algoritmo *Fuzzy C-Means* modificado.

3.2.1 Classificação cruzada (Acc)

Indica o quanto um teste de classificação binária identifica corretamente ou exclui uma condição (verdadeiros positivos/verdadeiros negativos).

$$Acc.\left(U^1, U^2\right) = \max_{\varsigma \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^{c} \left(n_{00}^{(i,\varsigma(i))} + n_{11}^{(i,\varsigma(i))} \right) \tag{3.2}$$

3.2.2 Medida de Comparação F1

Média harmônica entre precisão e recuperação (RIJSBERGEN, 1979).

$$F_1 = \frac{\pi\rho}{\left(\frac{\pi+\rho}{2}\right)} = \frac{2\pi\rho}{\pi+\rho} \tag{3.3}$$

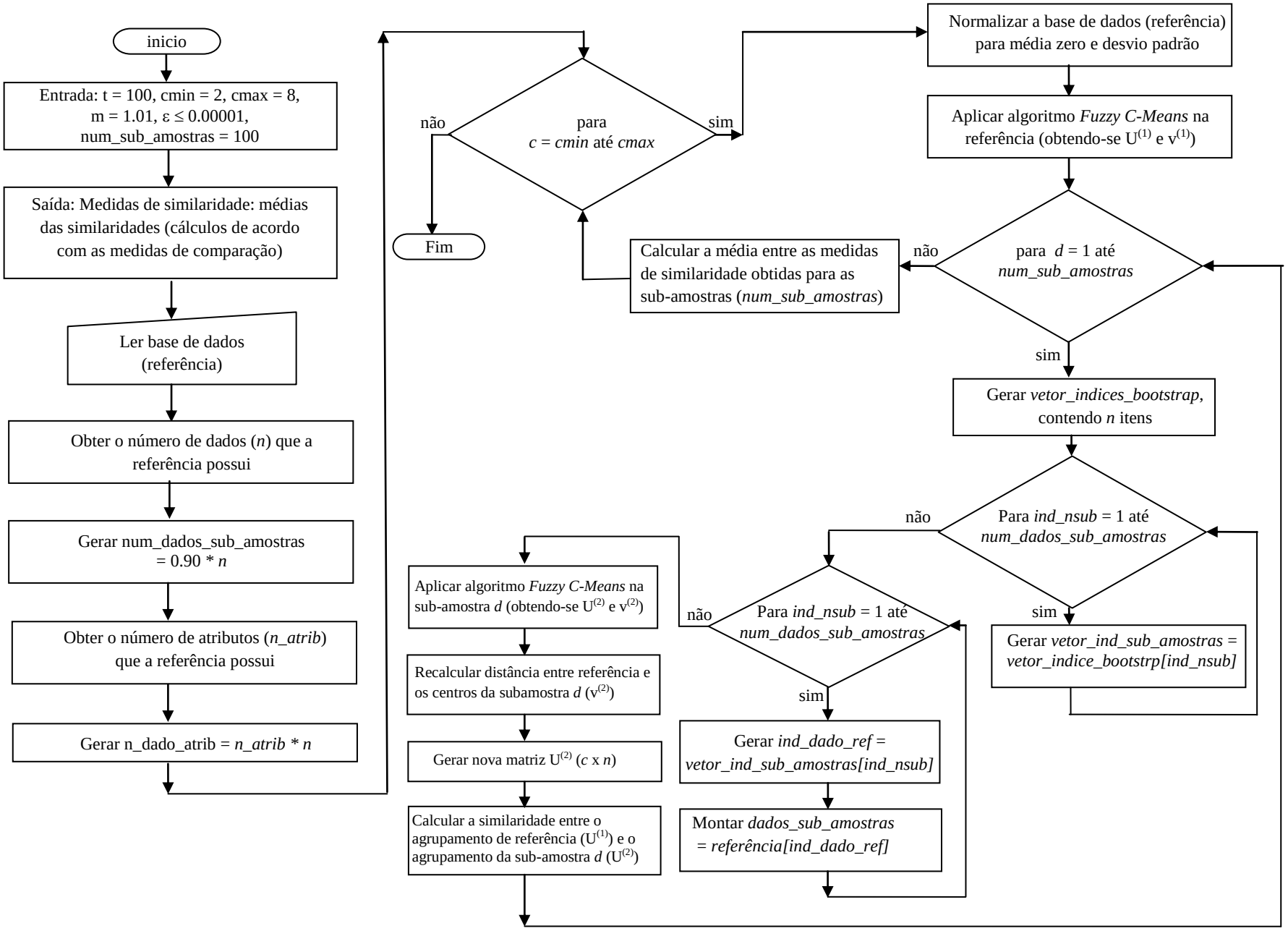
Onde:

- π representa precisão; e,
- ρ representa recuperação.

A permutação de todas as linhas da matriz é a melhor forma de se encontrar a linha da matriz $U^{(1)}$ que mais se iguale a linha da matriz de partição $U^{(2)}$, calcula-se precisão e recuperação por:

$$\pi = \max \varsigma \in \prod(c) \frac{1}{c} \sum_{i=1}^{c} \frac{n_{11}^{(i,\varsigma(i))}}{n_{01}^{(i,\varsigma(i))} + n_{11}^{(i,\varsigma(i))}} \tag{3.4}$$

Figura 3.7 - Fluxograma para cálculo do número de grupos em bases de dados.



$$\rho = \max \varsigma \in \prod(c) \frac{1}{c} \sum_{i=1}^{c} \frac{n_{11}^{(i,\varsigma(i))}}{n_{10}^{(i,\varsigma(i))} + n_{11}^{(i,\varsigma(i))}} \tag{3.5}$$

tem-se:

$$F_1\left(U^1, U^2\right) = \max \varsigma \in \prod(c) \frac{1}{c} \sum_{i=1}^{c} \frac{2\pi_i \varsigma(i) \rho_{i,} \varsigma(i)}{\pi_i \varsigma(i) + \rho_{i,} \varsigma(i)} \tag{3.6}$$

Onde: Π(c) representa o conjunto de todas as permutação de linhas *c* da matriz de partição.

3.2.3 Hubert (Hub)

Corrige estatísticas ao acaso. Pode ser interpretado como a raiz quadrada da media qui-quadrado. Valores altos indicam alta probabilidade de aceitar a partição (BORGELT, 2006).

$$Hubert = \frac{N_{00}N_{11} - N_{10}N_{01}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})(N_{01} + N_{00})(N_{10} + N_{00})}} \tag{3.7}$$

3.2.4 Jaccard

Esta medida calcula a probabilidade de que dados pertencentes ao mesmo grupo em uma das partições também pertençam ao mesmo grupo em outra partição (BOUTIN; HASCOëT, 2004).

$$Jaccard = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{3.8}$$

3.2.5 Índice Randômico (Rand)

Calcula a proporção de concordância entre as partições (BOUTIN; HASCOëT, 2004; JAIN; DUBES, 1988).

$$Rand = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}} \tag{3.9}$$

3.2.6 Fowlkes and Mallows (Fowlkes)

Indica a probabilidade de que determinados dados pertençam a um mesmo grupo na partição $\mathbf{U}^{(1)}$ se eles fazem parte de um mesmo grupo na partição $\mathbf{U}^{(2)}$ (FOWLKES&MALLOWS, 1983).

$$Fowlkes-Mallows = \sqrt{\underbrace{\frac{N_{11}}{N_{11}+N_{10}}}_{z}.\underbrace{\frac{N_{11}}{N_{11}+N_{01}}}_{g}} \tag{3.10}$$

Onde:

- z indica a probabilidade de que determinados dados pertençam a um mesmo grupo na partição $\mathbf{U}^{(1)}$ se eles fazem parte de um mesmo grupo na partição;
- g é a probabilidade de que dados pertençam ao mesmo grupo em $\mathbf{U}^{(2)}$ se eles pertencem ao mesmo grupo em $\mathbf{U}^{(1)}$.

Na sequência deste trabalho, serão apresentados os resultados encontrados na simulação geradas pelo algoritmo *Fuzzy c-Means* variando o índice de nebulosidade.

CAPÍTULO 4

4 RESULTADOS E DISCUSSÃO

Diversos estudos têm sido propostos para a validação de partições de dados produzidos pelo algoritmo de agrupamento *Fuzzy c-Means*. E dentre esses estudos tem-se o importante papel do índice de nebulosidade, também conhecido como índice de fuzificação ou expoente de ponderação, que para PAL e BEZDEK (1995) é muito sensível tanto para valores baixos quanto para valores elevados.

O objetivo deste trabalho foi variar o índice de nebulosidade para encontrar a melhor faixa de valores a serem utilizadas para a classificação dos dados e das medidas utilizadas, consequentemente obtenção de melhores particionamentos. A qualidade de comparação foi baseada em medidas de comparação tradicionais: Classificação Cruzada (Acc), F1, Hubert (Hub), Jaccard, Índice Randômico (Rand) e *Fowlkes & Mallows*. As bases de dados utilizadas foram três bases Artificiais, *Iris, Wine e Brady 1*.

As Tabelas 4.1 a 4.6 apresentam os resultados gerados nas simulações com as bases de dados Artificial 1, Artificial 2, Artificial 3, *Iris, Wine e Brady 1*, com índice de nebulosidade m = [1,01 1,02 1,03 1,04 1,05 1,06 1,07 1,1 1,2 1,5 1,8 2,0 2,3 2,5 3,0 3,5 4,0 4,5] para todas as medidas de comparação e número de grupos variando de 2 a 8.

Todas as medidas foram calculadas usando *t-norma produto* e *nil-potent mínimo*. Os valores em negrito nas tabelas 4.1 a 4.6, correspondem aos valores máximos de cada medida. Nas Figuras 4.1, 4.2, 4.3 e 4.5, as linhas tracejadas delimitam a faixa em que as medidas apresentam as respostas corretas para os números de grupos do conjunto de dados.

O valor final para cada medida foi obtido aplicando-se a média entre os resultados das 100 sub-amostras. O valor máximo obtido para cada medida, entre todos os particionamentos realizados, indica o número de grupos definidos pela medida como sendo o mais correto (mais estável) para a base de dados em análise.

Observa-se na Tabela 4.1 e na Figura 4.1 (base de dados Artificial 1) que a melhor faixa de variação do expoente m para todas os índices está no intervalo de *1,01* a *2,0*. Para os índices de comparação Hub, Jaccard e *Fowlkes* esta faixa se estendeu, começando em *1,02* a *3,5*.

Tabela 4.1 - Simulação base de dados Artificial 1

Valor m	Acc	nº de grupos	F1	nº de grupos	Hub	nº de grupos	Jaccard	nº de grupos	Rand	nº de grupos	Fowlkes	nº de grupos
1,01	0,0000	0	0,0000	0	**0,9874**	**4**	**0,9812**	**4**	**0,9954**	**4**	**0,9905**	**4**
1,02	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,03	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,04	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,05	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,06	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,07	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,1	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**	**1,0000**	**4**
1,2	**1,0000**	**4**	**0,9999**	**4**	**0,9998**	**4**	**0,9997**	**4**	**0,9999**	**4**	**0,9998**	**4**
1,5	**0,9920**	**4**	**0,9839**	**4**	**0,9680**	**4**	**0,9529**	**4**	**0,9881**	**4**	**0,9759**	**4**
1,8	**0,9550**	**4**	**0,9098**	**4**	**0,8607**	**4**	**0,8089**	**4**	**0,9489**	**4**	**0,8944**	**4**
2,0	**0,9180**	**4**	**0,8359**	**4**	**0,7792**	**4**	**0,7101**	**4**	**0,9213**	**4**	**0,8305**	**4**
2,3	**0,8601**	**4**	0,7242	3	**0,6781**	**4**	**0,5953**	**4**	0,9018	8	**0,7463**	**4**
2,5	0,8328	8	0,6770	3	**0,6253**	**4**	**0,5376**	**4**	0,9107	8	**0,6993**	**4**
2,8	0,8176	8	0,6135	3	**0,5630**	**4**	**0,4712**	**4**	0,9364	8	**0,6406**	**4**
3,0	0,8132	8	0,5825	2	**0,5307**	**4**	**0,4372**	**4**	0,9596	8	**0,6084**	**4**
3,5	0,8012	8	0,5235	2	**0,4773**	**4**	**0,3770**	**4**	0,9872	8	**0,5476**	**4**
4,0	0,7929	8	0,5439	2	**0,4501**	**4**	0,3664	2	0,9973	8	0,5361	2
4,5	0,7906	8	0,5399	2	0,4397	3	0,3627	2	0,9993	8	0,5323	2

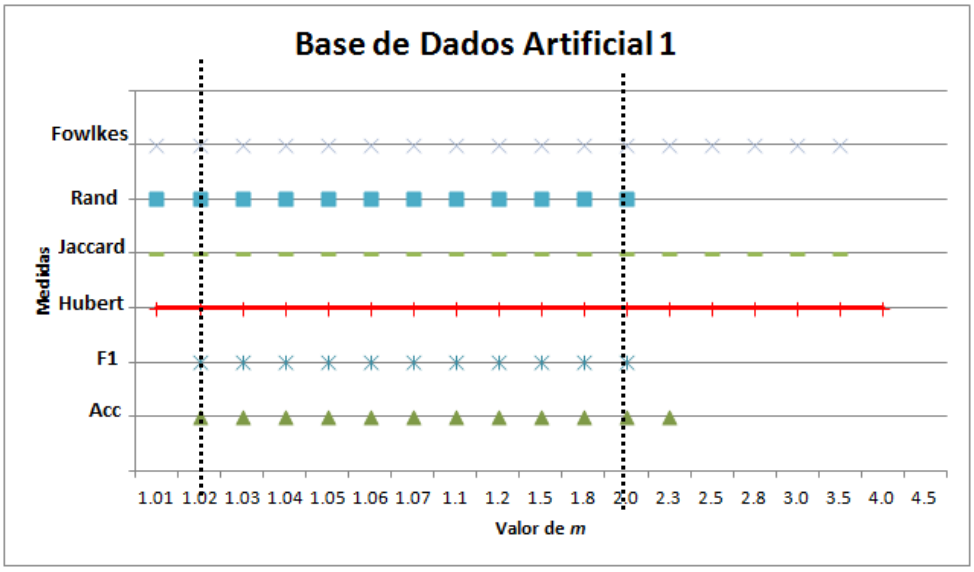


Figura 4.1 - Gráfico base de dados Artificial 1

A melhor faixa de *m* para todos os índices de comparação na base de dados Artificial 2 está entre *1,04* a *2,0*. Na Tabela 4.2, observa-se que para os índices de comparação Hubert (Hub), Jaccard e *Fowlkes*, a melhor a faixa de *m* está entre *1,01* a *2,8*.

Tabela 4.2 - Simulação base de dados Artificial 2

Valor m	Acc	nº de grupos	F1	nº de grupos	Hub	nº de grupos	Jaccard	nº de grupos	Rand	nº de grupos	Fowlkes	nº de grupos
1,01	0,0000		0,0000	0	**0,9818**	**7**	**0,9710**	**7**	**0,9955**	**7**	**0,9844**	**7**
1,02	0,9764	4	0,9586	2	**0,9862**	**7**	**0,9802**	**7**	**0,9964**	**7**	**0,9883**	**7**
1,03	0,9780	4	0,9593	4	**0,9417**	**7**	**0,9053**	**7**	**0,9868**	**7**	**0,9493**	**7**
1,04	**0,9941**	**7**	**0,9775**	**7**	**0,9836**	**7**	**0,9765**	**7**	**0,9957**	**7**	**0,9860**	**7**
1,05	**0,9949**	**7**	**0,9811**	**7**	**0,9886**	**7**	**0,9836**	**7**	**0,9970**	**7**	**0,9903**	**7**
1,06	**0,9938**	**7**	**0,9763**	**7**	**0,9820**	**7**	**0,9741**	**7**	**0,9953**	**7**	**0,9847**	**7**
1,07	**0,9981**	**7**	**0,9922**	**7**	**0,9929**	**7**	**0,9899**	**7**	**0,9982**	**7**	**0,9940**	**7**
1,1	**0,9951**	**7**	**0,9812**	**7**	**0,9862**	**7**	**0,9802**	**7**	**0,9964**	**7**	**0,9882**	**7**
1,2	**0,0000**	**7**	**0,9921**	**7**	**0,9924**	**7**	**0,9890**	**7**	**0,9980**	**7**	**0,9935**	**7**
1,5	**0,9893**	**7**	**0,9610**	**7**	**0,9514**	**7**	**0,9246**	**7**	**0,9881**	**7**	**0,9583**	**7**
1,8	**0,9654**	**7**	**0,8784**	**7**	**0,8500**	**7**	**0,7727**	**7**	**0,9650**	**7**	**0,8703**	**7**
2,0	**0,9388**	**7**	**0,7847**	**7**	**0,7625**	**7**	**0,6581**	**7**	**0,9477**	**7**	**0,7924**	**7**
2,3	**0,8959**	**7**	0,6813	2	**0,6530**	**7**	**0,5282**	**7**	**0,9353**	**7**	**0,6903**	**7**
2,5	**0,8730**	**7**	0,6539	2	**0,6028**	**7**	**0,4717**	**7**	**0,9327**	**7**	**0,6407**	**7**
2,8	0,8430	8	0,6216	2	**0,5348**	**7**	**0,3996**	**7**	0,9400	8	**0,5708**	**7**
3,0	0,8322	8	0,6056	2	**0,4975**	**7**	0,3856	2	0,9495	8	0,5565	2
3,5	0,8097	8	0,5755	2	**0,4313**	**7**	0,3686	2	0,9786	8	0,5387	2
4,0	0,8010	8	0,5567	2	0,4130	4	0,3590	2	0,9947	8	0,5284	2
4,5	0,7949	8	0,5437	2	0,4160	4	0,3526	2	0,9990	8	0,5214	2

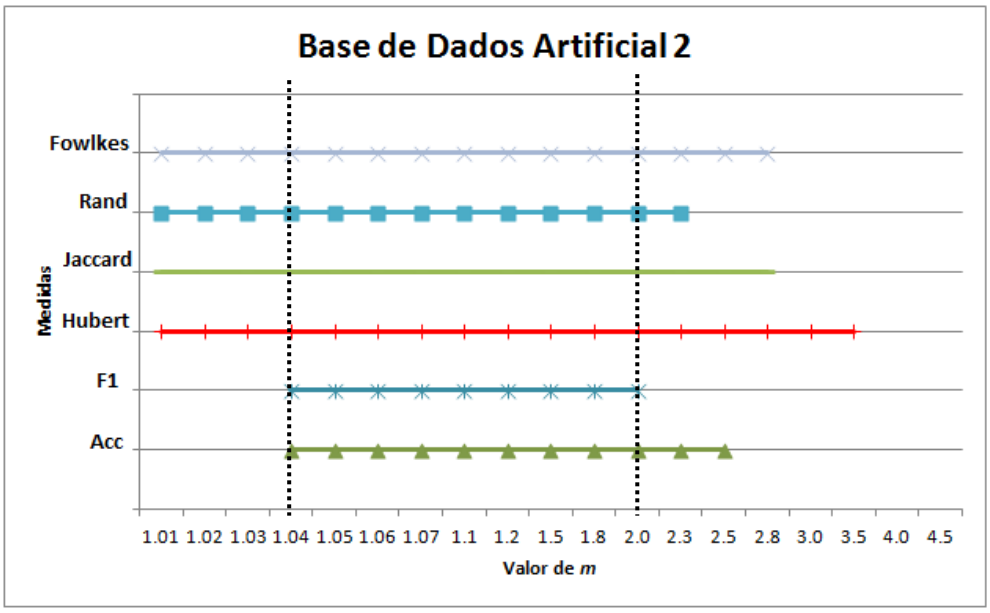


Figura 4.2 - Base de dados Artificial 2 - melhor valor de *m*

Para a base de dados Artificial 3, a melhor faixa de *m* está entre *1,03* a *1,2*. Nota-se que o índice de comparação Hubert (Hub) estende sua faixa até o valor *3,0*.

Tabela 4.3 - Simulação base de dados Artificial 3

Valor m	Acc	nº de grupos	F1	nº de grupos	Hub	nº de grupos	Jaccard	nº de grupos	Rand	nº de grupos	Fowlkes	nº de grupos
1,01	0,0000		0,0000	0	**0,9853**	**5**	**0,9780**	**5**	**0,9952**	**5**	**0,9883**	**5**
1,02	**0,9972**	**5**	0,9929	3	**0,9845**	**5**	**0,9764**	**5**	**0,9950**	**5**	**0,9876**	**5**
1,03	**0,9978**	**5**	**0,9944**	**5**	**0,9860**	**5**	**0,9778**	**5**	**0,9955**	**5**	**0,9888**	**5**
1,04	**0,9975**	**5**	**0,9939**	**5**	**0,9845**	**5**	**0,9755**	**5**	**0,9951**	**5**	**0,9876**	**5**
1,05	**0,9971**	**5**	**0,9928**	**5**	**0,9820**	**5**	**0,9716**	**5**	**0,9942**	**5**	**0,9856**	**5**
1,06	**0,9957**	**5**	**0,9888**	**5**	**0,9757**	**5**	**0,9629**	**5**	**0,9922**	**5**	**0,9806**	**5**
1,07	**0,9961**	**5**	**0,9904**	**5**	**0,9759**	**5**	**0,9622**	**5**	**0,9923**	**5**	**0,9807**	**5**
1,1	**0,9946**	**5**	**0,9866**	**5**	**0,9667**	**5**	**0,9481**	**5**	**0,9894**	**5**	**0,9734**	**5**
1,2	**0,9880**	**5**	**0,9702**	**5**	**0,9288**	**5**	**0,8921**	**5**	**0,9773**	**5**	**0,9430**	**5**
1,5	**0,9477**	**5**	0,8866	3	**0,7637**	**5**	0,6989	3	**0,9257**	**5**	0,8228	3
1,8	**0,8856**	**5**	0,7710	3	**0,6139**	**5**	0,5568	3	0,9069	8	0,7153	3
2,0	0,8445	8	0,7187	2	**0,5502**	**5**	0,4910	3	0,9243	8	0,6586	3
2,3	0,8243	8	0,6631	2	**0,5044**	**5**	0,4302	2	0,9535	8	0,6016	3
2,5	0,8163	8	0,6351	2	**0,4852**	**5**	0,4110	2	0,9721	8	0,5826	2
2,8	0,8046	8	0,6043	2	**0,4634**	**5**	0,3913	2	0,9875	8	0,5625	2
3,0	0,8018	8	0,5882	2	**0,4547**	**5**	0,3816	2	0,9942	8	0,5524	2
3,5	0,7942	8	0,5614	2	0,4296	4	0,3663	2	0,9992	8	0,5362	2
4,0	0,7892	8	0,5448	2	0,4246	3	0,3572	2	0,9999	8	0,5264	2
4,5	0,7873	8	0,5338	2	0,4315	3	0,3515	2	1,0000	8	0,5201	2

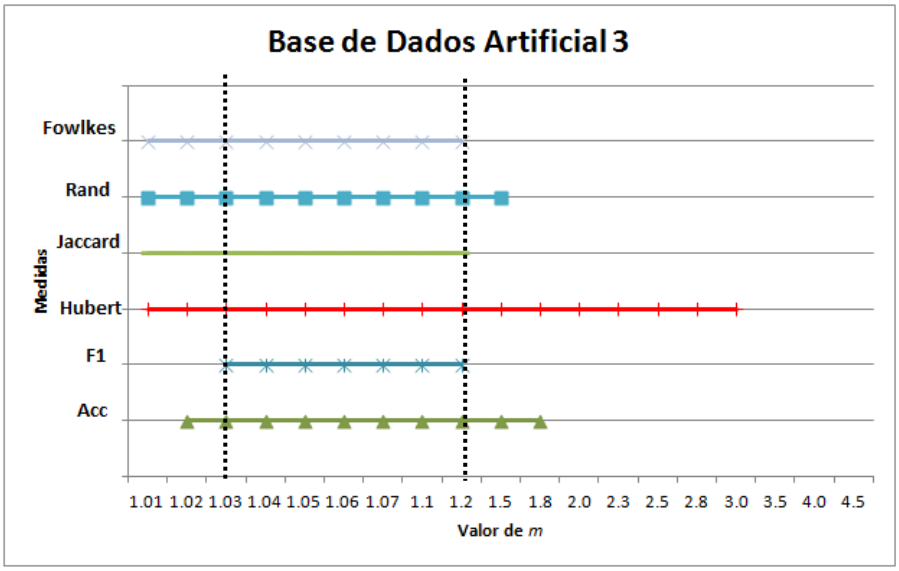


Figura 4.3 - Base de dados Artificial 3 - melhor valor de *m*

A base de dados *Iris* apresentou valor correto de agrupamento (3 grupos) somente para o Índice de Comparação Hubert (Hub) a partir da faixa de *2,3* a *4,5*.

Tabela 4.4- Simulação base de dados *Iris*

Valor m	Acc	nº de grupos	F1	nº de grupos	Hub	nº de grupos	Jaccard	nº de grupos	Rand	nº de grupos	Fowlkes	nº de grupos
1,01	0,9965	2	0,9961	2	0,9859	2	0,9875	2	0,9930	2	0,9936	2
1,02	0,9971	2	0,9968	2	0,9886	2	0,9899	2	0,9943	2	0,9948	2
1,03	0,9958	2	0,9953	2	0,9831	2	0,9851	2	0,9916	2	0,9924	2
1,04	0,9957	2	0,9952	2	0,9829	2	0,9849	2	0,9915	2	0,9923	2
1,05	0,9944	2	0,9938	2	0,9777	2	0,9802	2	0,9889	2	0,9899	2
1,06	0,9945	2	0,9938	2	0,9779	2	0,9803	2	0,9890	2	0,9900	2
1,07	0,9939	2	0,9932	2	0,9758	2	0,9785	2	0,9880	2	0,9891	2
1,1	0,9924	2	0,9915	2	0,9697	2	0,9730	2	0,9850	2	0,9863	2
1,2	0,9868	2	0,9853	2	0,9476	2	0,9536	2	0,9740	2	0,9762	2
1,5	0,9459	2	0,9412	2	0,8101	2	0,8390	2	0,9056	2	0,9124	2
1,8	0,8787	2	0,8716	2	0,6293	2	0,7036	2	0,9063	8	0,8260	2
2,0	0,8465	8	0,8241	2	0,5213	2	0,6303	2	0,9197	8	0,7732	2
2,3	0,8240	8	0,7613	2	**0,4317**	**3**	0,5504	2	0,9505	8	0,7100	2
2,5	0,8153	8	0,7260	2	**0,4093**	**3**	0,5118	2	0,9648	8	0,6771	2
2,8	0,8061	8	0,6832	2	**0,3881**	**3**	0,4700	2	0,9781	8	0,6395	2
3,0	0,8016	8	0,6599	2	**0,3870**	**3**	0,4495	2	0,9859	8	0,6202	2
3,5	0,7938	8	0,6170	2	**0,3926**	**3**	0,4145	2	0,9970	8	0,5861	2
4,0	0,7896	8	0,5881	2	**0,3976**	**3**	0,3932	2	0,9999	8	0,5645	2
4,5	0,7871	8	0,5683	2	**0,4075**	**3**	0,3793	2	1,0000	8	0,5500	2

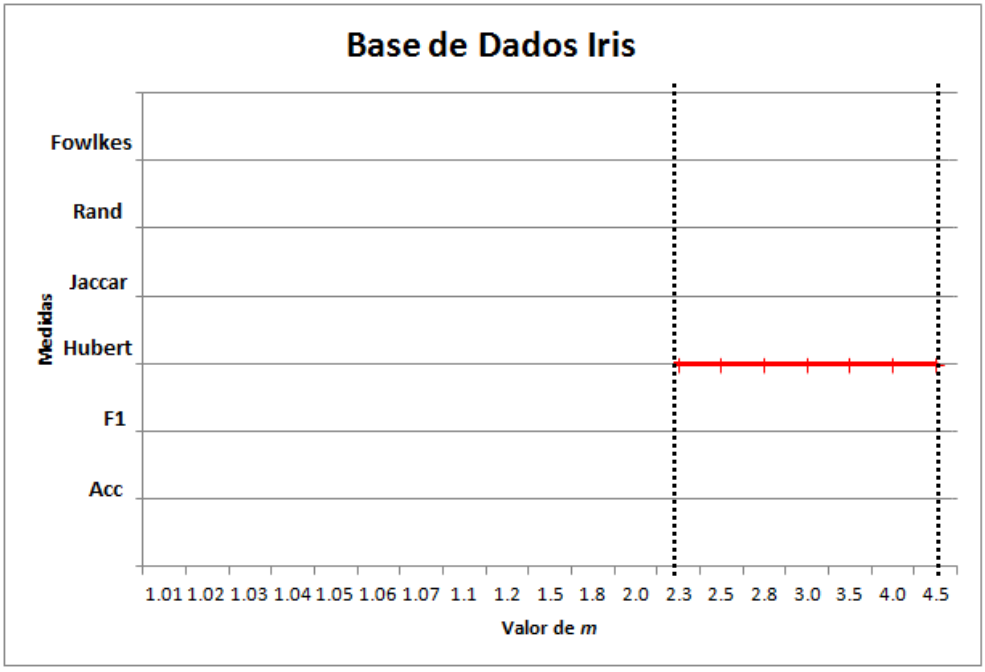


Figura 4.4 - Base de dados *Iris* - melhor valor de *m*

A base de dados *Wine* teve a melhor de faixa de *m* entre *1,01* a *1,2* para todos os índices de comparação. Observa-se que para o Índice de Hubert (Hub) teve valor ótimo em todos os valores de testados.

Tabela 4.5 - Simulação base de dados *Wine*

Valor m	Acc	nº de grupos	F1	nº de grupos	Hub	nº de grupos	Jaccard	nº de grupos	Rand	nº de grupos	Fowlkes	nº de grupos
1,01	**0,9847**	**3**	**0,9768**	**3**	**0,9331**	**3**	**0,9186**	**3**	**0,9698**	**3**	**0,9560**	**3**
1,02	**0,9841**	**3**	**0,9755**	**3**	**0,9286**	**3**	**0,9124**	**3**	**0,9674**	**3**	**0,9537**	**3**
1,03	**0,9811**	**3**	**0,9705**	**3**	**0,9177**	**3**	**0,9004**	**3**	**0,9625**	**3**	**0,9464**	**3**
1,04	**0,9759**	**3**	**0,9635**	**3**	**0,8993**	**3**	**0,8791**	**3**	**0,9541**	**3**	**0,9345**	**3**
1,05	**0,9794**	**3**	**0,9683**	**3**	**0,9086**	**3**	**0,8879**	**3**	**0,9584**	**3**	**0,9404**	**3**
1,06	**0,9774**	**3**	**0,9648**	**3**	**0,9008**	**3**	**0,8799**	**3**	**0,9546**	**3**	**0,9357**	**3**
1,07	**0,9764**	**3**	**0,9645**	**3**	**0,9009**	**3**	**0,8816**	**3**	**0,9547**	**3**	**0,9356**	**3**
1,1	**0,9759**	**3**	**0,9635**	**3**	**0,8931**	**3**	**0,8697**	**3**	**0,9516**	**3**	**0,9301**	**3**
1,2	**0,9651**	**3**	**0,9468**	**3**	**0,8479**	**3**	**0,8205**	**3**	**0,9306**	**3**	**0,9013**	**3**
1,5	**0,9173**	**3**	**0,8738**	**3**	**0,6946**	**3**	**0,6675**	**3**	0,8950	7	**0,8005**	**3**
1,8	0,8723	8	**0,7585**	**3**	**0,5267**	**3**	**0,5200**	**3**	0,9005	8	**0,6842**	**3**
2,0	0,8567	8	0,7169	2	**0,4548**	**3**	0,4792	2	0,9194	8	0,6476	2
2,3	0,8331	8	0,6486	2	**0,4033**	**3**	0,4198	2	0,9495	8	0,5910	2
2,5	0,8230	8	0,6375	2	**0,3886**	**3**	0,4152	2	0,9676	8	0,5867	2
2,8	0,8137	8	0,5886	2	**0,3784**	**3**	0,3813	2	0,9851	8	0,5519	2
3,0	0,8094	8	0,5804	2	**0,3751**	**3**	0,3748	2	0,9946	8	0,5452	2
3,5	0,8033	8	0,5566	2	**0,3866**	**3**	0,3627	2	0,9991	8	0,5323	2
4,0	0,8003	8	0,5427	2	**0,3933**	**3**	0,3550	2	0,9999	8	0,5240	2
4,5	0,7975	8	0,5332	2	**0,4040**	**3**	0,3501	2	1,0000	8	0,5186	2

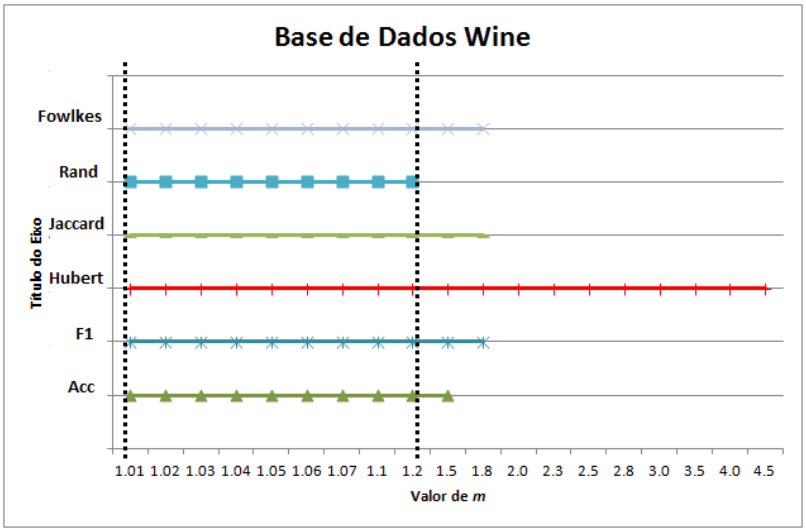


Figura 4.5 - Base de dados *Wine* - melhor valor de *m*

O grupo correto da base de dados *Brady1* é de 5 ou 6 grupos, e as simulação apresentaram esse número de grupos somente para o índice Hubert (Hub) nos valores de *m* de *1,03* à *2,0*, conforme demonstrado na Tabela 4.6 e na figura 4.6.

Tabela 4.6 - Simulação base de dados *Brady1*

Valor m	Acc	nº de grupos	F1	nº de grupos	Hub	nº de grupos	Jaccard	nº de grupos	Rand	nº de grupos	Fowlkes	nº de grupos
1,01	0,0000	0	0,0000	0	0,0000	0	0,0000	0	1,0000	2	0,0000	0
1,02	0,0000	0	0,0000	0	0,0000	0	0,0000	0	1,0000	2	0,0000	0
1,03	0,8913	2	0,8883	2	**0,6490**	**5**	0,7093	2	0,9175	8	0,8232	2
1,04	0,8864	2	0,8835	2	**0,6516**	**6**	0,6977	2	0,9109	8	0,8153	2
1,05	0,8910	2	0,8887	2	**0,6544**	**6**	0,7066	2	0,9104	6	0,8210	2
1,06	0,8831	2	0,8803	2	**0,6543**	**6**	0,6865	2	0,9135	8	0,8090	2
1,07	0,8771	2	0,8755	2	**0,6121**	**6**	0,6686	2	0,9100	8	0,7967	2
1,1	0,8586	2	0,8548	2	**0,5874**	**6**	0,6393	2	0,9134	8	0,7773	2
1,2	0,8430	8	0,7707	2	**0,5127**	**6**	0,5129	2	0,9175	8	0,6774	2
1,5	0,8160	8	0,6326	2	**0,3901**	**5**	0,3951	2	0,9446	8	0,5664	2
1,8	0,7993	8	0,5682	2	**0,3789**	**5**	0,3641	2	0,9736	8	0,5339	2
2,0	0,7931	8	0,5475	2	**0,3704**	**5**	0,3549	2	0,9869	8	0,5239	2
2,3	0,7884	8	0,5335	2	0,3355	4	0,3477	2	0,9968	8	0,5160	2
2,5	0,7862	8	0,5256	2	0,3332	3	0,3444	2	0,9989	8	0,5124	2
2,8	0,7845	8	0,5172	2	0,3297	3	0,3420	2	0,9997	8	0,5097	2
3,0	0,7836	8	0,5142	2	0,0102	2	0,3405	2	0,9999	8	0,5081	2
3,5	0,7827	8	0,5093	2	0,0067	2	0,3384	2	1,0000	7	0,5057	2
4,0	0,7822	8	0,5066	2	0,0048	2	0,3376	2	1,0000	8	0,5048	2
4,5	0,7820	8	0,5047	2	0,0034	2	0,3370	2	1,0000	4	0,5041	2

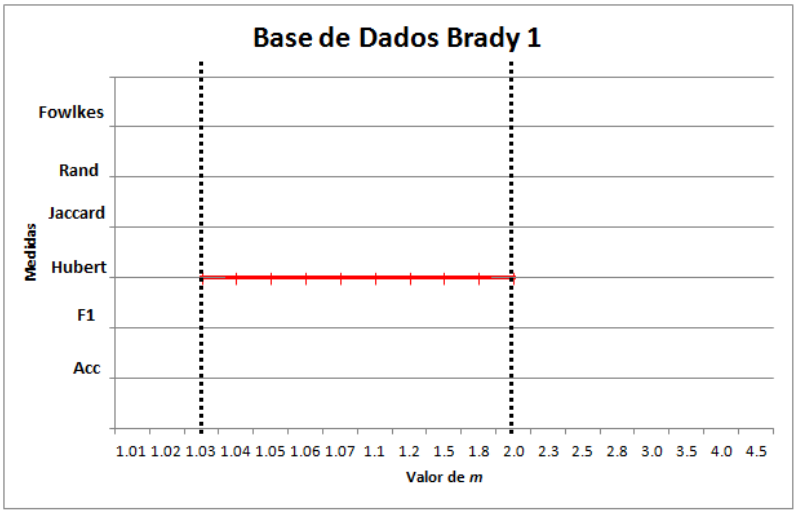


Figura 4.6 - Base de dados *Brady 1* - melhor valor de *m*

A Tabela 4.7 apresenta um sumário das simulações feitas neste trabalho, onde os pontos representam os valores máximos que coincidiram com o número correto de grupos de cada base de dados em todos os índices de Comparação: Classificação Cruzada (Acc), F1, Hubert (Hub), Jaccard, Randômico e *Fowlkes & Mallows*. Pode-se observar que a melhor faixa de valor de *m* para todos as bases de dados é entre *1.04* e *1.2* exceto para base de dados *Iris* e *Brady 1*, pois, tiveram os valores corretos de agrupamento somente para o índice de comparação Hubert (Hub) tendo as faixas *1,04* a *1,2* e *2,5* a *4,5* respectivamente.

Tabela 4.7 - Sumário dos resultados das simulações

valor de *m*	Artificial 1	Artificial 2	Artificial 3	*Iris*	*Wine*	*Brady* 1
1,01	●				●	
1,02	●				●	
1,03	●		●		●	
1,04	●	●	●		●	● *
1,05	●	●	●		●	● *
1,06	●	●	●		●	● *
1,07	●	●	●		●	● *
1,1	●	●	●		●	● *
1,2	●	●	●		●	● *
1,5	●	●				● *
1,8	●	●				● *
2,0	●	●				● *
2,3				● *		
2,5				● *		
2,8				● *		
3,0				● *		
3,5				● *		
4,0				● *		
4,5				● *		

* os valores de *m* para base de dados *Iris* e *Brady1* somente coincidiu com o número correto de grupos no Índice Hubert (Hub)

CAPÍTULO 5

5 CONCLUSÃO

Este trabalho apresentou um estudo sobre a variação do índice de nebulosidade, utilizando reamostragem *bootstrapping* e o algoritmo *Fuzzy C-Means*, que leve aos agrupamentos mais estáveis.

Utilizou-se seis bases de dados sendo três bases de dados Artificiais as bases reais *Iris*, *Wine* e *Brady 1*. Observa-se que para grupos bem definidos e separados linearmente como o caso das bases de dados Artificial 1, Artificial 2 e Artificial 3 predomina a faixa de valores do índice de nebulosidade entre *1,04* e *2,0* nas seis medidas de comparação Acc, F1, Hubert, Jaccard, Rand e *Fowlkes*. Para a Base de Dados Iris nota-se que de *2,3* a *4,5* é a faixa ideal mantendo os grupos, porém, somente para o índice de comparação Hubert, isso considerando que somente um grupo é linearmente separável. Já na base de dados *Wine*, a variação do índice nebuloso apresentou variação de *1,01* a *1,2* para todas as seis medidas de comparação. Na base de dados *Brady 1* a melhor faixa foi de *1,03* até *2,0* somente para o índice de comparação Hubert.

Importante ressaltar que o índice *Hubert* manteve em todas as bases de dados o número correto de grupos para os valores de *m.*

PAL e BEZDEK (1995) apresentaram um estudo semelhante usando o algoritmo *Fuzzy c-Means*, a base de dados *Iris* e uma base de dados denominada "Normal4" (semelhante com as Artificiais 1, 2 e 3) e seus resultados sugerem que a melhor faixa do valor de *m* é a faixa de *1,5* a *2,5* e cujo ponto médio utilizado por outros usuários é *m* igual *2,0*. Os resultados apresentados neste trabalho, variam daqueles apresentados por PAL e BEZDEK, essa variação reafirma a importância da determinação da influência do valor de *m* nos resultados de agrupamentos, pois a diferença encontrada pode ser devida aos índices de medida adotados e as bases de dados que foram utilizadas, indicando que o valor de *m* é dependente desses parâmetros.

Para trabalhos futuros sugere-se:

- Analisar outras bases de dados e que também possuam maior quantidade de grupos;
- Análise com outras medidas para determinação daquelas que sejam menos sensíveis à variação do valor de *m*.

CAPÍTULO 6

6 REFERÊNCIAS

BALKE, C. L.; MERZ, C. J. UCI **Repository of Machine Learning Databases**. Irvine. CA. USA: University of California. 1998.
http://mlearn.ics.uci.edu/MLRepository.html [Online. acesso em 29/05/2012].

BOHRER, T. R. J.; HUNGRIA. M. **Avaliação de cultivares de soja quanto à fixação biológica do Nitrogênio**. Dissertação (Mestrado) Universidade Estadual de Londrina. PR. Brasil, 1997.

BORGELT, C. **Resampling for fuzzy clustering**. *In: Proc. Symposium on Fuzzy Systems in Computer Science*. Otto-von-Guericke-Universitat. Magdeburg. Germany, 2006.

BORGELT, C. **Prototype-based Classification and Clustering**. Tese (Doutorado) - Otto-von-Guericke-Universitat. Magdeburg. Germany, November 2005.

CHRIST, R. E. **Classificação de Bactérias do gênero Bradyrhizobium usando uma rede neural ART2 com dados de eletroforese de genes ribossomais**. Dissertação (Mestrado) Universidade de São Paulo. São Carlos, SP. Brasil, 2007.

CROWLEY, P. H. **Resampling methods for computation-intensive data analysis in ecology and evolution**. Annual Review of Ecology and Systematics, v. 23, p. 405-447, 1992.

EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the Boosttrap**. Chapman and Hall. New York, 1993.

GERMANO, M. G.; MENNA, P.; MOSTASSO, F. L.; HUNGRIA. M. **Rflp analysis of the rrna operon of a brazilian collection of bradyrhizobial strains from 33 legume species**. International Journal of Systematic and Evolutionary Microbiology, The Society for General Microbiology. v. 56. p. 217-229. 2006.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS. M. **Cluster validity cheching methods**: Part ii. ACM SIGMOD. ACM Press. New York. NY, USA. v. 31. n. 3. p. 19-27. September 2002.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS. M. **Cluster validity methods**: Part i. ACM SIGMOD. ACM Press. New York. NY. USA. v. 31. n. 2. p. 40-45. September 2002.

HUBERT, L.; ARABIE, P. **Comparing partitions**. Journal of Classification, v. 2, p. 193-218, 1985.

HUNGRIA, M.; VARGAS, M. A. T.; ARAÚJO, R. S. **Fixação biológica do nitrogênio em feijoeiro**. Biologia dos Solos dos Cerrados. EMBRAPA-CPAC, p. 189-225. 1997.

HÖPPNER, F.; KLAWONN, F.; KRUSE, R.; RUNKLER. T. **Fuzzy Cluster Analysis: Methods for Classification. Data Analysis and Image Recognition.** Chichester. England: John Wiley and Sons, 1999.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data clustering: A review**. ACM Computing Surveys (CSUR). ACM Press. vol. 31. no. 3. pp. 264-323. New York, September 1999.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Upper Saddle River. NJ. USA. Prentice Hall, 1988.

JOHNSON, R. A. **Applied Multivariate Statistical Analysis**. Prentice Hall. 3rd. ed. New Jersey, USA: Prentice Hall, 1992.

LAW, M. H.; JAIN, H. A. K. **Cluster Validity by Bootstrapping Partitions**. East Lansing. Michigan, USA, 2003.

LLERENA, S. E. **Mapeamento de Dados Genômicos Utilizando Multidimensional Scaling**. Dissertação (Mestrado) | Universidade de São Paulo. São Carlos. SP. Brasil, 2008.

MATHWORKS, I. S. M. T. **The Language of Technical Computing**. Natick. Massachusetts. USA: [s.n.]. 1998.

MENGER, K. Statistical metrics. **National Academy of Science of the United States of America**. v. 28. n. 12. p. 535-537. December 1942.

MILAGRE, S. T. **Análise do Número de Grupos em Bases de Dados Incompletas Utilizando Agrupamentos Nebulosos e Reamostragem Bootstrap**. Tese Doutorado USP São Carlos. 2008.

MILAGRE. S. T. **Análise de Estabilidade de Cluster em uma Coleção Brasileira de Bactérias Dizotróficas do Gênero Bradtyrhizobium.** Dissertação (Mestrado). Universidade Estadual de Londrina. PR. 2003.

NUOVO, A. G. D.; CATANIA, V.; NUOVO, S. D.; BUONO, S. **Envolving fuzzy c-means: An intelligent technique for efficient diagnosis of children mental retardation level from databases with missing values**. In: International Conference on Artificial Intelligence (ICAI´06). Las Vegas. USA: [s.n]. 2006.

PAL, N. R.; BEZDEK, J. C. **On cluster validity for the fuzzy c-means model**. IEEE Transactions on Fuzzy Systems. *IEEE Computer Society*. Washington. DC. USA. vol. 3. no. 3. pp. 370-379. August 1995.

ROSSUM, D. van; SCHUURMANS, F. P.; GILLIS, M.; MUYOTCHA, A.; VERSELD, H. W. Van. **Genetic and phenotypic analyses of Bradyrhizobium strains nodulating peanuts (Arachis hypogea L.)**. Applied and Environmental Microbiology. v. 61. p. 1599-1609. 1995.

ROTH, V.; LANGE, T.; BRAUN, M. L.; BUHMANN, J. M. **A resampling approach tocluster validation. In: 15th Computational Statistics (COMPSTAT***)*. Physica-Verlag Heidelberg. Germany: Berlin 2002.

SARKAR, M.; LEONG, T. Y. **Fuzzy k-means clustering with missing values**. In: American Medical Informatics Association Annual Symposium (AMIA). Medical Publishers. pp. 588-59. Philadelphia, 2001.

SANDRI, S.; CORREA, C. **Lógica nebulosa**. In: V Escola de Redes Neurais. ITA, São José dos Campos, SP, Brasil: [s.n.], 1999.

SO, R. B.; LADHA, J. K.; YOUNG, J. P. W. **Photosynthetic symbionts of Aeschynomene spp. form a cluster with bradyrhizobia on the basis of fatty acid and rrna analysis**. International Journal of Systematic Bacteriologyy. v. 44. p. 392-403. 1994.

TALAVERA, E. R. V. **Métodos Bayesianos Aplicados em Taxonomia Molecular**. Dissertação (Mestrado) | Universidade de São Paulo, São Carlos. SP. Brasil, 2007.

TIMM, H.; DÖRING, C.; KRUSE, R. **Diferent approaches to fuzzy clustering of incomplete datasets**. Elsevier - International Journal of Approximate Reasoning. v. 35. n. 3. p. 239-249. March 2004.

XIE, X. L.; Beni, G. **A validity measure for fuzzy clustering**. IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE Computer Society. vol. 13. no. 8. pp. 841-847. Washington, November 1991.

XU, R.; WUNSCH II, D. C. **Survey of clustering algorithms**. IEEE Transactions on Neural Networks. vol. 16. no. 3. pp. 645-678, 2005.

ZADEH, L. A. **Fuzzy logic. neural networks and soft computing. Comunications of the ACM.** ACM Press. vol. 37. no. 3. pp. 77-84. New York. USA. March 1994.

ZHANG, D.; CHEN, S. **Clustering incomplete data using kernel-based fuzzy c-means algorithm**. Neural Processing Letters. Kluwer Academic. vol. 18. no. 12. pp. 155-162, Netherlands, NY. USA. December 2003.

7 ANEXO A

Extraído da Tese de Selma Terezinha Milagre (autorizado pela autora)

3.5 Medindo a Validade do Agrupamento

O objetivo de *agrupar* é descobrir, automaticamente, o intrínseco agrupamento de um conjunto de dados (LAW; JAIN, 2003), assim, a principal preocupação em um processo de agrupamento é revelar a organização dos padrões dentro de grupos *sensível*, os quais nos permitem descobrir similaridades e diferenças, assim como derivar inferências úteis sobre eles (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a).

Diferentes algoritmos de agrupamentos já foram propostos para diferentes aplicações e tamanhos de bases de dados (JAIN; MURTY; FLYNN, 1999; LAW; JAIN, 2003; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a). A aplicação de um algoritmo a uma base de dados tem como objetivo, assumindo-se que a base de dados oferece uma tendência ao agrupamento, descobrir suas partições naturais. Entretanto, o processo de agrupamento é considerado um processo não supervisionado, tendo em vista que não são pré-definidas classes e nem exemplos que mostrem que tipo de relações desejáveis devem ser válidas entre os dados. Desta forma, os vários algoritmos de agrupamento baseiam-se em algumas suposições para definir o particionamento de uma base de dados, o que pode levar a resultados diferentes dependendo das características do conjunto de dados, como por exemplo geometria e densidade de distribuição dos grupos, e dos valores dos parâmetros de entrada (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a).

Assim, adicionado às diferenças ente os resultados dos algoritmos, deve-se considerar que os algoritmos de agrupamentos produzem um modelo de particionamento para uma base de dados, quer existam ou não, tornando-se necessária a validação de tais agrupamentos, seja verificando se o modelo de agrupamento obtido é o que mais se adéqua ao conjunto de dados ou avaliando-se a qualidade do agrupamento.

O procedimento de avaliação que pode fornecer uma resposta quantitativa para estas questões é denominado *validade do agrupamento* (*cluster validity*), o qual é o objetivo de muitos esforços de pesquisadores (LAW; JAIN, 2003; PAL; BEZDEK, 1995; XIE; BENI, 1991; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a; HALKIDI, BATISTAKIS; VAZIRGIANNIS, 2002b).

Em termos gerais existem três métodos para investigar a validade de um agrupamento que são os critérios de avaliação externos, internos e relativos. Nos critérios de avaliação externos a avaliação dos resultados do algoritmo de agrupamento baseia-se na comparação entre o particionamento obtido pelo algoritmo com uma estrutura de grupos pré-especificada que é imposta ao conjunto de dados e reflete nossa intuição sobre a estrutura que a mesma deve possuir. Nos critérios internos, o objetivo é avaliar o agrupamento resultante, utilizando-se somente características intrínsecas da base de dados, ou seja, baseiam-se em medidas de avaliação internas. No método de critérios relativos, a idéia básica é avaliar a estrutura do agrupamento por meio da comparação do mesmo com outros esquemas de agrupamentos resultantes do mesmo algoritmo, mas com diferentes valores para os parâmetros (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002b; LAW; JAIN, 2003)

A reamostragem com reposição (*bootstrapping* (EFRON; TIBSHIRANI, 1993)) pode ser usada para estimar índices realtivos, comparando-se o agrupamento resultante para uma da amostra da base de dados e um agrupamento de referência. Assim, utiliza-se os métodos baseados em reamostragem para verificar onde a base de dados possui uma estrutura de agrupamentos ou onde o resultado é um artefato do algoritmo e também para selecionar-se o modelo do número de grupos (LAW; JAIN, 2003; BORGELT, 2005). Neste trabalho utilizou-se o método interno de comparação combinado com o critério relativo reamostragem *bootstrap*.

3.5.1 Comparando Matrizes de Partição

Medidas de avaliação relativas para grupos comparam duas matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$, onde $\mathbf{U}$ é uma matriz ($c \times n$) e $\mathbf{U} = (\mu_{ij})$. Lembrando que: c é o número de grupos e n é o número de dados contidos na base de dados, e ainda, ($1 \leq i \leq c$) e ($1 \leq j \leq n$). As matrizes de partição a serem comparadas podem ser, por exemplo, uma matriz obtida pelo algoritmo de agrupamento e a outra uma matriz de referência (modelo). Faz-se a comparação diretamente, aos pares, ou seja, compara-se cada par de linhas, sendo uma de cada matriz de partição (índice i refere-se à $\mathbf{U}^{(1)}$ e o índice r refere-se à $\mathbf{U}^{(2)}$). Formalmente, utiliza-se uma tabela de contingência 2 x 2 (Tabela 3.3) para cada par de linhas, uma de cada matriz de partição.

Tabela 3.3 – Tabela de contingência para comparar duas linhas de duas matrizes de partição *crisp* (Adaptado de BORGELT, 2005)).

	$\mu_{(rj)}^{(2)} = 1$	$\mu_{(rj)}^{(2)} = 0$	$\sum$
$\mu_{(ij)}^{(1)} = 1$	$n_{(11)}^{(i,r)}$	$n_{(10)}^{(i,r)}$	$n_{1.}^{(i,r)}$
$\mu_{(ij)}^{(1)} = 0$	$n_{(01)}^{(i,r)}$	$n_{(00)}^{(i,r)}$	$n_{0.}^{(i,r)}$
$\sum$	$n_{.1}^{(i,r)}$	$n_{.0}^{(i,r)}$	n

Assim, para cada par de índices de grupo, $(i,r) \in \{1, ..., c\}^2$ e cada par $(a,b) \in \{0,1\}^2$, calcula-se:

$$n_{ab}^{(i,r)}(U^{(1)},U^{(2)}) = \sum_{j=1}^{n}\left((1-a)+(2a-1)\mu_{ij}^{(1)}\right)\cdot\left((1-b)+(2b-1)\mu_{rj}^{(2)}\right) \tag{3.14}$$

Substituindo-se os valores de *a* e *b*, tem-se:

$$n_{11}^{(i,r)}(U^{(1)},U^{(2)}) = \sum_{j=1}^{n}\mu_{ij}^{(1)}\cdot\mu_{rj}^{(2)} \tag{3.15}$$

$$n_{01}^{(i,r)}(U^{(1)},U^{(2)}) = \sum_{j=1}^{n}\left(1-\mu_{ij}^{(1)}\right)\cdot\mu_{rj}^{(2)} \tag{3.16}$$

$$n_{10}^{(i,r)}(U^{(1)},U^{(2)}) = \sum_{j=1}^{n}\mu_{ij}^{(1)}\cdot(1-\mu_{rj}^{(2)}) \tag{3.17}$$

$$n_{00}^{(i,r)}(U^{(1)},U^{(2)}) = \sum_{j=1}^{n}(1-\mu_{ij}^{(1)})\cdot(1-\mu_{rj}^{(2)}) \tag{3.18}$$

Sendo que o termo n_{11} representa os verdadeiros positivos, n_{01} os falso positivo, n_{10} os falso negativos, e n_{00} os verdadeiros negativos. Por exemplo, no agrupamento rígido (ou *crisp*) n_{11} representa o número de dados que estão no *i-ésimo* grupo da partição $\mathbf{U}^{(1)}$ e no *r-ésimo* grupo da partição $\mathbf{U}^{(2)}$. Onde a operação booleana 'e' é formalmente expressa pelo produto nas Equações 3.14 a 3.18.

No caso de agrupamentos nebulosos ($\mu_{\cdot} \in [0,1]$) e considerando-se que o produto expressa a operação booleana 'e' (*and)*, pode-se utilizar o conceito de *t-norma* (Seção 3.3.2) para calcular-se os termos definidos nas Equações 3.15 a 3.18.

Por exemplo, reescrevendo-se a Equação 3.14, temos:

$$n_{ab}^{(i,r)}(U^{(1)},U^{(2)}) = \sum_{j=1}^{n}\left(\underbrace{(1-a)+(2a-1)\mu_{ij}^{(1)}}_{x}\right)\cdot\left(\underbrace{(1-b)+(2b-1)\mu_{rj}^{(2)}}_{y}\right)$$

Desta forma, considerando-se o produto entre *x* e *y* tem-se a *t-norma* $\mathsf{T}(x,y)$. Aplicando-se a *t-norma* Zadeh (Tabela 3.2) na Equação 3.16, tem-se:

$$x = (1-\mu_{(ij)}^{(1)}) \qquad \text{e} \qquad y = (1-\mu_{(ij)}^{(2)})$$

Assim:

$$\mathrm{T}_{\min}(x,y) = \min((1-\mu_{(ij)}^{(1)}),\mu_{(rj)}^{(2)} \tag{3.19}$$

(Note-se que mostra-se aqui um dos termos referentes ao somatório).

Calculados os números $n_{(11)}^{(i,r)}, n_{(10)}^{(i,r)}, n_{(01)}^{(i,r)}, n_{(00)}^{(i,r)}$, pode-se utilizá-los no cálculo de medidas de comparação binária, tais como F_1 (RIJSBERGEN, 1979) e classificação cruzada (ou *accuracy*). Quanto maior o valor dessas duas medidas, mais similares são as matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$.

Antes de definir a medida F_1, faz-se necessário o detalhamento de duas medidas, as quais são necessárias no cálculo de F_1, que são *precisão* e *recuperação*, definidas em (RIJSBERGEN, 1979). Originalmente, essas medidas são calculadas para cada grupo e descrevem o quão bem um grupo é identificado pelo método classificador. A comparação é feita entre o grupo real e o predito pelo classificador (problema dois grupos) (BORGELT, 2005).

Para calcular *precisão* e *recuperação*, faz-se uma tabela de contingência 2 x 2 (Tabela 3.4) para cada grupo, baseadas nos grupos verdadeiros *v* associados com os dados x_j, com $1 \leq j \leq n$, e os grupos *o* que são preditos pelo classificador. Os elementos para da tabela de contingência é qualquer grupo *r* ($\forall r : 1 \leq r \leq c$)

Tabela 3.4 – Tabela de contingência para o cálculo de *precisão* e *recuperação* (Adaptado de (BORGELT, 2005)).

	$o = r$	$o \neq r$	$\sum$
$v = r$	$n_{(11)}^{(r)}$	$n_{(10)}^{(r)}$	$n_{1.}^{(r)}$
$v \neq r$	$n_{(01)}^{(r)}$	$n_{(00)}^{(r)}$	$n_{0.}^{(r)}$
$\sum$	$n_{.1}^{(r)}$	$n_{.0}^{(r)}$	n

Assim, dado um grupo *r*, a *precisão* representa a taxa de verdadeiros positivos para todos os dados classificados como grupo *r*, ou seja, a fração de dados que o classificador agrupo corretamente:

$$\pi_r = \frac{n_{11}^{(r)}}{n_{11}^{(r)} + n_{01}^{(r)}} \tag{3.20}$$

A *recuperação* representa a taxa de verdadeiros positivos para todos os dados que atualmente pertencem ao grupo *r*, isto é, é a fração de dados do grupo *r* que são identificados pelo classificador:

$$\rho_r = \frac{n_{11}^{(r)}}{n_{11}^{(r)} + n_{10}^{(r)}} \tag{3.21}$$

Quando deseja-se comparar vários grupos (em toda a base de dados), é preciso calcular a média das medidas individuais. Pode-se utilizar, por exemplo a *média-macro* (macro-averaging) (BORGELT, 2005):

$$\pi_{macro} = \frac{1}{c}\sum_{r=1}^{c} \frac{n_{11}^{(r)}}{n_{01}^{(r)} + n_{11}^{(r)}} \tag{3.22}$$

$$\rho_{macro} = \frac{1}{c}\sum_{r=1}^{c} \frac{n_{11}^{(r)}}{n_{10}^{(r)} + n_{11}^{(r)}} \tag{3.23}$$

- A medida de comparação binária F_1:

A medida F_1 é a harmônica entre *precisão* e *recuperação* (RIJSBERGEN, 1979):

$$F_1 = \frac{\pi\rho}{\left(\frac{\pi+\rho}{2}\right)} = \frac{2\pi\rho}{\pi+\rho} \tag{3.24}$$

Quando deseja-se encontrar a linha da matriz de partição $\mathbf{U}^{(1)}$ que mais se iguale a uma linha da matriz de partição $\mathbf{U}^{(2)}$, a melhor forma é utilizar-se permutação de todas as linhas da matriz ($\mathbf{U}$ é uma matriz *c x n*), ou seja, a permutação de todos os grupos, assim, calcula-se *precisão* e *recuperação* por:

$$\pi = \max_{\zeta\in\Pi(c)} \frac{1}{c}\sum_{i=1}^{c} \frac{n_{11}^{(i,\zeta(i))}}{n_{01}^{(i,\zeta(i))} + n_{11}^{(i,\zeta(i))}} \tag{3.25}$$

$$\rho = \max_{\zeta\in\Pi(c)} \frac{1}{c}\sum_{i=1}^{c} \frac{n_{11}^{(i,\zeta(i))}}{n_{10}^{(i,\zeta(i))} + n_{11}^{(i,\zeta(i))}} \tag{3.26}$$

Assim, tem-se:

$$F_1(U^{(1)},U^{(2)}) = \max_{\zeta\in\Pi(c)} \frac{1}{c}\sum_{i=1}^{c} \frac{2\pi_{i,\zeta(i)}\rho_{i,\zeta(i)}}{\pi_{i,\zeta(i)} + \rho_{i,\zeta(i)}} \tag{3.27}$$

Onde *Π(c)* representa o conjunto de todas as permutações de linhas *c* da matriz de partição.

Por exemplo, para um agrupamento em cinco grupos (*c=5*), teríamos a comparação do grupo c = 1 da matriz de partição $\mathbf{U}^{(1)}$ com os grupos *c = 1, 2, 3, 4 5* da matriz de partição $\mathbf{U}^{(2)}$). A seguir compara-se o grupo *c = 2* da matriz de partição $\mathbf{U}^{(1)}$ com os grupos *c = 1, 2, 3, 4, 5* da matriz de partição $\mathbf{U}^{(2)}$), e assim por diante até que se compare o último grupo de $\mathbf{U}^{(1)}$ (no caso *c = 5*) com todos os grupos de $\mathbf{U}^{(2)}$.

- A medida de comparação binária classificação cruzada (Acc):

Indica o quanto um teste de classificação binária identifica corretamente ou exclui uma condição, ou seja, é a proporção de resultados verdadeiros (verdadeiros positivos e verdadeiros negativos) em relação às matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$.

$$Acc.(U^{(1)},U^{(2)}) = \max_{\zeta\in\Pi(c)} \frac{1}{cn}\sum_{i=1}^{c}\left(n_{00}^{(i,\zeta(i))} + n_{11}^{(i,\zeta(i))}\right) \tag{3.28}$$

3.5.2 Comparando Matrizes de Coincidência

Matriz de coincidência é uma matriz *n x n*, e indica para cada par de dados onde esses dados estão em um mesmo grupo ou não. Calcula-se uma matriz de coincidência Ψ a partir de uma matriz de partição **U**, sendo $\Psi = (\psi_{jl})$, onde $1 \leq j, l \leq n$ (BORGELT, 2005):

$$\psi_{jl} = \sum_{i=1}^{c} \mu_{ij} \cdot \mu_{il} \tag{3.30}$$

Após calcular-se $\Psi^{(1)}$ e $\Psi^{(2)}$, pode-se gerar os números que comparam as coincidências na distribuição dos dados dentro dos grupos. Sendo cada par $(a,b) \in \{0,1\}^2$, tem-se:

$$N_{ab}\left(\Psi^{(1)}, \Psi^{(2)}\right) = \sum_{j=2}^{n} \sum_{l=1}^{j-1} \left((1-a) + (2a-1)\vartheta_{jl}^{(1)}\right) \cdot \left((1-b) + (2b-1)\vartheta_{jl}^{(2)}\right) \tag{3.31}$$

Substituindo-se os valores de a e b, tem-se:

$$N_{11}\left(\Psi^{(1)}, \Psi^{(2)}\right) = \sum_{j=2}^{n} \sum_{l=1}^{j-1} \psi_{jl}^{(1)} \cdot \psi_{jl}^{(2)} \tag{3.32}$$

$$N_{10}\left(\Psi^{(1)}, \Psi^{(2)}\right) = \sum_{j=2}^{n} \sum_{l=1}^{j-1} \psi_{jl}^{(1)} \cdot (1 - \psi_{jl}^{(2)}) \tag{3.33}$$

$$N_{01}\left(\Psi^{(1)}, \Psi^{(2)}\right) = \sum_{j=2}^{n} \sum_{l=1}^{j-1} (1 - \psi_{jl}^{(1)}) \cdot \psi_{jl}^{(2)} \tag{3.34}$$

$$N_{00}\left(\Psi^{(1)}, \Psi^{(2)}\right) = \sum_{j=2}^{n} \sum_{l=1}^{j-1} (1 - \psi_{jl}^{(1)}) \cdot (1 - \psi_{jl}^{(2)}) \tag{3.35}$$

Sendo que N_{11} representa o número de pares de dados que estão em um mesmo grupo em ambas as partições, N_{10} e N_{01} representam o número de pares que estão em um mesmo grupo em uma partição mas em grupos diferentes na outra partição e, N_{00} indica o número de pares de dados que estão grupos difernetes em ambas as partições.

Da mesma forma que citou-se na Seção 3.5.1, o produto representado na Equação 3.3.1 também pode ser substituído por uma *t-norma*, quando trabalha-se com conjuntos nebulosos.

Note-se que como essa substituição pela *t-norma* pode ser feita tanto na Equação 3.30 quanto na Equação 3.31, quando trabalha-se com matrizes de coincidência pode-se faze várias combinações entre a primeira e segunda *t-normas*, ou seja, pode-se utilizar na Equação 3.30 a *t-norma* produto e na Equação 3.31 também a *t-norma* produto, ou utilizar-se *t-norma* mínimo e *t-norma* produto nas Equações 3.30 e 3.31, respectivamente, e assim por diante.

De posse dos números N_{11}, N_{01}, N_{10} e N_{00} pode-se calcular várias medidas de comparação, tais como:

- **Coeficiente e Jaccard**

Define a similaridade entre duas partições ($\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$). Esse índice calcula a probabilidade de que dados pertencentes ao mesmo grupo em uma das partições também pertençam ao mesmo grupo na outra partição (BOUTIN; HASCOëT, 2004), ignorando a informação negativa N_{00}, que representa os pares que são atribuídos para grupos diferentes em ambas as partições (BORGELT, 2005).

$$Jaccard = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{3.36}$$

- **Índice Randômico**

Este índice calcula a proporção de concordância entre as partições. Assim, calcula a probabilidade de que os dados que pertençam a um grupo na partição $\mathbf{U}^{(1)}$ também pertençam ao mesmo grupo na partição $\mathbf{U}^{(2)}$ ou, que pertençam a grupos diferentes (BOUTIN HASCOëT, 2004; JAIN; DUBES, 1988).

$$Rand = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}} \tag{3.37}$$

- **Índice de *Fowlkes & Mallows***

Foi proposto pro *Fowlkes e Mallows* (1983):

$$Fowlkes-Mallows = \sqrt{\underbrace{\frac{N_{11}}{N_{11}+N_{10}}}_{x} \cdot \underbrace{\frac{N_{11}}{N_{11}+N_{01}}}_{y}} \tag{3.38}$$

Onde o termo *x* indica a probabilidade de que determinados dados pertençam a um mesmo grupo na partição $\mathbf{U}^{(1)}$ se eles fazem parte de um mesmo grupo com $\mathbf{U}^{(2)}$. O terno *y* é a probabilidade de que dados pertençam ao mesmo grupo em $\mathbf{U}^{(2)}$ se eles pertencem ao mesmo grupo em $\mathbf{U}^{(1)}$ (BOUTIN; HASCOëT, 2004.

Reescrevendo Equação 3.38.

$$Fowlkes-Mallows = \frac{N_{11}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{011})}} \tag{3.39}$$

- **Índice Hubert**

Foi proposto por (HUBERT; ARABIIE, 1995). É uma normalização do índice randômico para corrigir a estatística que ocorre *ao acaso*. Desta forma, o valor máximo é 1 quando é alcançada uma perfeita concordância e 0 (se o valor das partições são selecionadas aleatoriamente (BOUTIN; HASCOëT, 2004; JAIN; DUBES, 1988). É também chamado de estatística Phi.

Pode ser interpretado como a raiz quadrada da medida qui-quadrado (χ^2). Valores altos indicam alta probabilidade de aceitar a partição (BORGELT, 2006).

$$Hubert = \frac{N_{00}N_{11} - N_{10}N_{01}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})(N_{01}+N_{00})(N_{10}+N_{00})}} \tag{3.40}$$

- **Índice Randômico Ajustado**

Frequentemente utiliza-se medidas que são correções de medidas existentes. Uma correção bastante comum é a normalização do índice correspondente para que ele apresente o valor *1* quando a partição casa perfeitamente com a partição real (ou modelo) e o valor *0* quando as partições são selecionadas ao acaso (JAIN; DUBES, 1988; GORDON, 1999).

$$Rand.Aj. = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{2(N_{11}N_{00}) + (N_{11}N_{00})(N_{01} + N_{10}) + N_{10}^2 + N_{01}^2} \quad (3.41)$$

Neste trabalho utilizou-se todas essas medidas na comparação entre matrizes de coincidência, sendo que para todas utilizou-se os valores máximos obtidos para a análise do número de grupos.

3.6 Reamostragem *Bootstrap*

Técnicas de reamostragem foram propostas há várias décadas, mas devido a serem métodos estatísticos computacionais intensivos, sua difusão só foi possível após o advento dos computadores modernos, com grande capacidade de processamento e custo relativamente baixo.

O procedimento *bootstrap* é uma técnica de reamostragem, bastante utilizada em diferentes situações estatísticas, que procura substituir a análise estatística teórica utilizando o poder da computação. O termo *bootstrap* surgiu da frase "*to pull oneself up one´s bootstrap*" retirado do texto *Adventures of Baron Munchausem* de Rudolpf Erick Raspe, século XVII, que dizia: "The Baron had fallen to the of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstrap". Descreve-se nesta parte do texto a cena em que o Barão Munchausem ao se ver afundando em um lago e sentindo que tudo estava pedido, acha que conseguirá emergir puxando os cadarços de seus próprios sapatos (EFRON; TIBSHIRANI, 1993).

A base desta técnica é a obtenção de um *novo* conjunto de dados por reamostragem do conjunto de dados original (EFRON; TIBSHIRANI, 1993). Assim, explica-se a metáfora *bootstrap*, que refere-se ao fato de que a base de dados é usada em sua própria análise estatística, de modo que, todo resultado obtido depende diretamente da base de dados original.

Os métodos de reamostragem se diferenciam, principalmente, pela forma em que as amostras são extraídas do conjunto de dados, que pode ser com reposição ou sem reposição. Na amostragem com reposição obtém-se um dado da base de dados para montar a amostra e então coloca-se o mesmo de volta para se usado posteriormente. Assim, uma mesma amostra pode conter dados repetidos. Na amostragem sem reposição, uma vez que obtém-se um determinado dado para compor a amostra, este torna-se indisponível, ou seja não se pode mais utilizá-lo para compor esta amostra.

A inferência estatística procura estabelecer as propriedades da população, partindo-se de uma amostra aleatória retirada da mesma. Assim, o método *bootstrap* obtém sua amostra via amostragem com reposição, para revelar algum padrão estrutural presente na base de dados. A ideia básica é de que os dados em si, vistos como distribuição de frequências, representam a melhor imagem disponível da distribuição de frequências da qual eles são amostrados (CROWLEY, 1992).

Seja uma amostra aleatória de tamanha *n* obtida de uma distribuição de probabilidades *F*:

$$F \rightarrow (x_1 x_2, x_3, ..., x_n) \tag{3.42}$$

A função de distribuição empírica $\widehat{F}$ (o símbolo chapéu indica quantidades calculadas a partir dos dados observados), é definida como sendo a distribuição discreta que atribui a probabilidade *1/n* para cada valor observado x_j, onde *j* = 1, 2, 3, ..., *n* (EFRON; TIBSHIRANI, 1993).

O método *bootstrap* depende da noção de amostra *bootstrap*. Uma amostra *bootstratp* é definida com sendo uma amostra aleatória de tamanho *n* retirada de *F*:

$$\widehat{F} \rightarrow (x_1^*, x_2^*, x_3^*, ..., x_n^*) \tag{3.43}$$

Assim,

$$x^* = (x_1^*, x_2^*, ..., x_n^*) \tag{3.44}$$

Desta forma, os dados *bootstrap* $x_1^*, x_2^*, ..., x_n^*$ formam uma amostra aleatória de tamanho *n* retirada da base de dados de *n* objetos. A notação utilizando-se o asterisco significa que x^* não é um novo conjunto de dados, mas apenas uma versão reamostrada de x, onde os dados podem aparecer zero vezes, uma vez, duas vezes, etc. Assim, pode-se ter, por exemplo: $x^* = (x_2, x_4, x_5, x_2, ..., x_4)$, onde $x_1^* = x_2, x_2^* = x_4, x_3^* = x_5, x_4^* = x_2$ *e* $x_n^* = x_4$ (EFRON; TIBSHIRANI, 1993).

O método *bootstrap* é baseado no princípio *plug-in*, um método simples de estimar parâmetros a partir de amostras. Seja um parâmetro θ obtido aplicando-se algum procedimento de avaliação numérica *t(.)* da função de distribuição *F*:

$$\theta = t(F) \tag{3.45}$$

A estimativa *plug-in* de um parâmetro $\theta = t(F)$ é definida como sendo:

$$\widehat{\theta} = t(\widehat{F}) \tag{3.46}$$

Assim, por este princípio, estima-se a função $\theta = t(F)$ da distribuição de probabilidade F pela mesma função da distribuição empírica $\widehat{F}$, $\widehat{\theta} = t(\widehat{F})$. A vantagem deste método está em produzir-se tendências e erros-padrões de forma automática, não importando a complexidade de θ (EFRON; TIBSHIRANI, 1993).